Breast density measurement for personalised screening

Katharina Holland

Breast density measurement for personalised screening

Katharina Holland

The research described in this thesis was carried out at the Diagnostic Image Analysis Group, Radboud University Medical Center (Nijmegen, The Netherlands). This work was funded by the EU FP7 project ASSURE.
Financial support for publication of this thesis was kindly provided by the department of Radiology and Nuclear Medicine of the Radboud University Medical Center, Nijmegen, The Netherlands.

Cover design: Promotie In Zicht, Arnhem.
Printed by Ipskamp Printing, Enschede.

ISBN: 978-94-028-0719-6

Breast density measurement for personalised screening

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op woensdag 4 oktober 2017
om 16.30 uur precies

door

Katharina Holland

geboren op 4 februari 1988
te Goch, Duitsland

Promotor:	**Prof. dr. ir. N. Karssemeijer**
Copromotoren:	**Dr. R. M. Mann**
	Dr. C. H. van Gils
	Universitair Medisch Centrum Utrecht
Manuscriptcommissie:	**Prof. dr. A. L. M. Verbeek**
	Prof. dr. R. M. Pijnappel
	Universitair Medisch Centrum Utrecht
	Dr. M. Lillholm
	Københavns Universitet, Denemarken

Breast density measurement for personalised screening

Doctoral Thesis

to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.H.J.M. van Krieken,
according to the decision of the Council of Deans
to be defined in public on Wednesday, October 4, 2017
at 16.30 hours

by

Katharina Holland

born on February 4, 1988
in Goch, Germany

Supervisor:	**Prof. dr. ir. N. Karssemeijer**
Co-supervisor:	**Dr. R. M. Mann**
	Dr. C. H. van Gils
	University Medical Center Utrecht
Doctoral Thesis Committee:	**Prof. dr. A. L. M. Verbeek**
	Prof. dr. R. M. Pijnappel
	University Medical Center Utrecht
	Dr. M. Lillholm
	Copenhagen University, Denmark

TABLE OF CONTENTS

1

Introduction

Breast cancer is the most common cancer diagnosis in women. About 464.000 women were diagnosed with breast cancer in Europe in 2012 and about 131.000 women died because of breast cancer [1]. Early detection is crucial to reduce breast cancer mortality. Therefore, many countries have implemented breast cancer screening programs in which asymptomatic women are screened regularly to find the breast cancer before it becomes palpable and symptomatic.
Even though the number of breast cancer deaths is decreasing, the incidence of breast cancer is increasing. This can be explained by the ageing of the population [2, 3]. The breast cancer incidence rates are highest for women between 50-75 years of age. Gender and age are the most important risk factor for breast cancer and so far the only risk factors that are considered in population screening. Many other risk factors are known. There are genetic factors, reproduction and hormonal factors, dietary factors and socioeconomic factors; an overview is given in Table 1.1. The risk factor breast density is one of the hormonal factors, as breast density decreases with age and after the menopause. Breast density refers to the amount of lobes, ducts and epithelial and connective tissue, also called fibroglandular tissue, as compared to the breast volume. This thesis focuses on mammographic breast density and its potential to be used as stratification tool for personalised breast cancer screening.

1.1 Breast cancer

The breast consists of several lobes that are connected with the nipple through ducts. The lobes are surrounded by fatty tissue and consist of lobules which are used for milk production. Breast cancer can manifest itself in several ways. The classification into different subtypes is done by a pathologist, who evaluates the tissue samples under the microscope. The most common breast cancers are the invasive ductal carcinoma (IDC) and the invasive lobular carcinoma (ILC). Invasive cancers start growing in one type of tissue and spread then into the surrounding tissue. The lesion might be palpable and visible on the mammogram as a round, oval or irregular mass. Depending on the surrounding tissue, spiculations might be visible. The IDCs are a group of tumours that fail to exhibit sufficient characteristics to achieve classification as a specific histological type. The ILC on the other side is characterised by cells individually dispersed or arranged in single-file linear patterns [11]. Abnormal cells within the ducts or the lobes that did not invade into the surrounding tissue are called carcinoma in situ. Also here, a differentiation between ductal carcinoma in situ (DCIS) and lobular carcinoma in situ (LCIS) is possible. DCIS and LCIS are non-obligatory precursors of cancer. DCIS is much more common than LCIS. Not all in situ carcinomas evolve into an invasive cancer while most DCIS are intensively treated resulting in a loss of life quality without any survival benefit.

Risk factor	Comparison group	Relative risk (RR)
BRCA gene mutations [4]	Mutation carriers compared to non carriers	10-30
Family history of breast cancer [5]	First degree relatives with breast cancer compared with no family history	2.1
Age at first child [6]	Older than 35 years compared to younger than 20 years	1.4
Parity [7]	Change in risk with every birth	Decrease in RR of 7%
Breast feeding [7]	Change in risk with every year of breastfeeding	Decrease in RR of 4.3%
Age at menarche [8]	Change in risk with every year decrease	1.15
Age at menopause [8]	Change in risk with every year increase	1.03
Alcohol consumption [9]	More than 45 g/day compared to no consumption	1.46
Breast density [10]	More than 75% breast density compared with less than 5%	4.64

Table 1.1: Associations between risk factors and breast cancer.

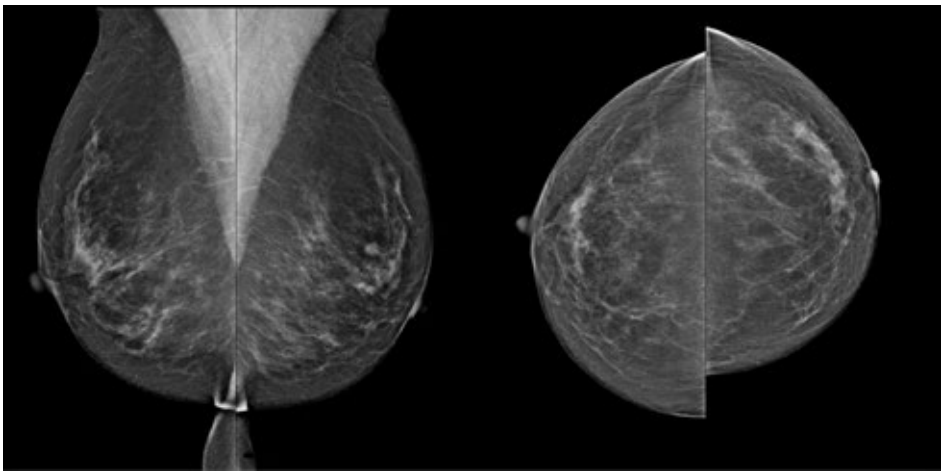


Figure 1.1: A set of mammograms, from the left to right: the right MLO view, the left MLO view, the right CC view and the left CC view.

1.2 Breast imaging

1.2.1 Mammography

Mammography is the most commonly used breast imaging technique. Mammograms are acquired within clinical practice and in breast cancer screening programs. To take a mammogram, the breast is positioned on the support plate and compressed with the compression paddle. Then, X-rays are sent through the breast. Previously, films were used to capture the image, but since the late 90s film mammography got replaced by full field digital mammography. Digital mammograms can be saved in two modes, as 'for processing' (raw) images and as 'for presentation' (processed) images. In raw images, the pixel values saved are proportional to the measured X-ray intensities. Each vendor has then its own algorithm to generate the 'for presentation' images based on the raw pixel data. The 'for presentation' images are used for evaluation by the radiologist. The X-ray attenuation is different for fat-involuted, fibroglandular and cancerous tissue. The difference in attenuation leads to the contrast between the different types of tissue on the mammogram. Fatty tissue is X-ray transparent and appears black on the mammogram, while fibroglandular tissue and cancerous tissue appear white in the image.
Usually, each breast is imaged in two different directions, the mediolateral oblique (MLO) view (angled side-view) and the cranio caudal (CC) view (top to bottom). If necessary, additional images are acquired with different angles, magnification views or spot views on the request of the radiologist. An exam consisting of the MLO and the CC images of the right and left breast is shown in Figure 1.1.

1.2.2 Digital Breast Tomosynthesis (DBT)

Mammography has its limitations; superpositions of fibroglandular tissue layers can obscure a lesion. On top of that, a superposition might look suspicious on the mammogram, leading to a referral. Digital Breast Tomosynthesis (DBT) has been developed to reduce the effect of overlapping tissue. As with mammography, the breast is compressed and X-rays are sent through the breast. The difference between mammography and DBT is the location of the X-ray source. The source of the X-rays is at a fixed position above the breast for mammography while the X-ray source is rotating over a limited range of angles over the breast for DBT. Several low dose X-ray images are acquired which are then used to generate a three-dimensional image. Instead of having only the projection of the breast into one image, several slices are reconstructed that prevent the effect of overlapping tissue. Each slice has a thickness of a few *mm* and the number of slices depends on the thickness of the compressed breast.

1.2.3 Magnetic Resonance Imaging (MRI)

With Magnetic Resonance Imaging (MRI) three-dimensional images are acquired. MRI has a high sensitivity. Therefore, it is used to screen women at a high risk for breast cancer (BRCA gene mutation carriers). These women are usually younger than the women participating in the population screening and they have a higher breast density. Additionally, MRI is used to locate known cancers within the breast and to estimate the size (volume) of the cancer. The disadvantage of MRI is, however, that it takes more time to obtain, read and evaluate the images. Furthermore, contrast agent administration is necessary, limiting the use of MRI in population screening.

1.2.4 Ultrasound

Breast ultrasound images can be divided into two different groups. First, there is hand held ultrasound. Images are acquired and evaluated at the same time. The examination is performed by a radiologist. Hand held ultrasound is used to follow up suspicious regions in the mammograms or to do an ultrasound guided biopsy. With three-dimensional whole breast ultrasound it is possible to generate an image of the entire breast that can be evaluated at a different point in time and that allows temporal comparison. Several volumes are obtained to cover the entire breast. The evaluation of whole breast ultrasound is very time consuming.

1.3 Breast cancer screening

Nowadays, most western countries have breast cancer screening programs. Starting at an age of 40-50 years women are regularly invited to get a mammogram. The interval between screening rounds varies between countries. The Netherlands has a screening interval of two

years. Initially women between 50-70 years of age were invited to participate, but given the increasing life expectancies, women are invited until the age of 75 since 1998.
Within the Dutch program, two radiologists review the mammograms independently. They refer the women for further investigations and/or imaging to the general practitioner or the hospital in case of a visible abnormality on the mammogram. Usually, ultrasound images are acquired and if a region looks suspicious on the ultrasound image, a biopsy is performed. The Netherlands has a rather low referral rate compared to other countries [12]. Women are referred in about 2.5% of the examinations. Most of the time, in about 72% of the referrals, no cancer is diagnosed. The cancer detection rate is at seven cancers per 1000 women screened [13].
Unfortunately, not all breast cancers are detected within screening programs, about 16-33% of the cancers is detected in between the screening rounds [14, 15]. They are called interval cancers. Compared to screen-detected cancers, interval cancers are detected in a later stage with a worse prognosis [16–18]. There are different reasons for a cancer detection outside the screening program. First, some cancers grow fast and develop from a small undetectable lesion at the time of screening to a palpable lesion in the screening interval. Second, the lesion might have been masked by the surrounding tissue and is therefore not distinguishable from normal tissue. Last, it is possible that the radiologist did not see the lesion or that the lesion was interpreted as normal tissue (observer error). Many interval cancers are visible on the screening mammogram in retrospect [19–21].

1.4 Breast density and masking

Studies have shown that women with high breast density have an up to six times increased risk for the development of breast cancer compared to women with low breast density [10, 22–24]. Breast density refers to the amount of fibroglandular tissue within the breast in comparison to the breast size. Breast density depends on hormonal factors and decreases with age and the menopause [25].
Fibroglandular tissue and cancerous tissue have the same attenuation for X-rays. They both appear white on the mammogram while fatty tissue is X-ray transparent and appears black. As a result, it is not always possible to differentiate fibroglandular tissue from cancerous tissue and the cancer remains hidden within the fibroglandular tissue structures. It is known that the sensitivity of mammography decreases with an increase in breast density [26–32]. Therefore, mammography is not the optimal screening modality for all women and screening programs taking into account breast density and other risk factors are under discussion [33, 34]. Especially in the United States, breast density is discussed in recent years. Many states passed breast density legislation, forcing radiologists to inform women about their breast density and the associated risks. Mammograms with increasing breast density from left to right are shown in Figure 1.2.

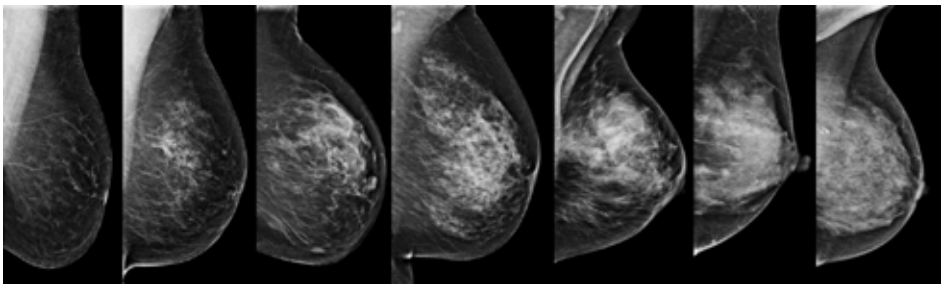


Figure 1.2: Mammograms with increasing breast density from the left to right.

The classification of breast density and breast density patterns is not new. Already in 1976, mammograms were categorised into different classes by Wolfe [35]. In 1982, Boyd and colleagues [36] introduced a scale based on mammographic density percentages. The classification was performed visually by a radiologist. Nowadays, the ACR BI-RADS density categories are commonly used in clinical practice. BI-RADS is a four-point categorisation. The 4th edition (2003) [37] included percentages, and the aim was to estimate the percentage dense area with respect to the area of the breast on the mammogram. The categories changed with the publication of the last (5th) edition (2013) [38]. Now, the aim is to describe the densest part in the mammogram. The masking effect is acknowledged with these changes. Even though the mammogram is not considered to belong to the highest density category with the 4th BI-RADS edition, there might be a region that is extremely dense and it is therefore considered to belong to the highest category with the 5th edition. Several studies have found considerable inter- and intra-reader variabilities when using the 4th [39–42] and 5th [43, 44] BI-RADS edition, respectively.

To use breast density as a stratification tool, it is necessary to have accurate and reliable breast density measurements. Furthermore, a continuous measurement might be preferable over a categorical to use breast density in risk models. Several algorithms have been developed in the past years to automate breast density assessments. Breast density measurements can be divided into two categories, depending on whether they aim to measure the volume of dense tissue or the area of the projected dense tissue.

Many algorithms have been developed to assign all pixels either to the class 'dense' or to the class 'non-dense'. Subsequently, the percentage dense area with respect to the breast area is determined. The disadvantage of these area based measurements is, that the 3D structure of the breast is not taken into account and that the measurement is not rotation invariant.

With the introduction of full field digital mammography volumetric breast density measurements became possible. Given the raw data, where the pixel intensity is proportional to the exposure, the fibroglandular tissue volume can be determined for each pixel location using an internal calibration [45–47]. Together with an estimation of the breast volume, the breast density measurement 'percent dense volume' is estimated by dividing

the fibroglandular tissue volume by the breast volume. By thresholding the percent density estimate, a categorical measure can be obtained that compares well to the BI-RADS density categories [48–50]. The internal reference method relies on the assumption that the chosen reference pixel value belongs to the projection of only fatty tissue. However, pixels representing fat only may not be present in dense breasts, causing an underestimation of density measurements [51–53]. An overview of breast density algorithms is given in He et al. [54]. Area based measurements with the semi automated program Cumulus [55] were considered gold standard for risk analysis for a long time. The advantage of this measurement is that it can be applied to all types of mammograms, film, and raw and processed digital mammograms. Since the transition to digital mammography, studies using volumetric breast density measurements are published as well. In the study by Eng et al. [24], breast cancer risk was estimated based on six breast density measurement techniques and the breast cancer risk association was strongest with measurements of volumetric breast density.

1.5 Alternatives to mammography and personalised screening

Mammography is the gold standard in population screening and most women are screened with mammography only. In the Netherlands, only women with a 50% lifetime risk of developing breast cancer (like BRCA gene mutation carriers, women at familial increased risk and those with with a prior history of radiation to the chest) have an adjusted screening scheme. Before introducing personalised screening and changing an entire screening program, it is necessary to show that the performance of the new scheme will be better than the old one. It is necessary to have a higher sensitivity, to find more cancers using the adjusted scheme than with the old one, and to have a comparable level of the positive predictive value, which is the ratio between justified referrals and overall referrals. Especially an increase in unjustified referrals should be avoided. Furthermore, it is necessary to keep in mind that the increase in cancer rate should not only be caused by less aggressive cancers that will never cause symptoms or death during the woman's expected lifetime (prevent increase of over-diagnosis). Last, the screening program should be cost efficient.
In the future, mammography might be replaced with DBT. To date, several (European) trials are ongoing that investigate the use of DBT in population screening. In the STORM trial [56] the combination of DBT and mammography is compared to mammography alone. The combination of DBT with mammography yielded a higher breast cancer detection rate than only mammography. Recalling based on both modalities separately (positive with mammography or DBT or positive with both modalities) increased the false positive rate compared to recalling based on mammography alone (5.5% vs. 2.0%) while recalling when an abnormality is visible with both modalities decreased the number of unjustified referrals (1.0%). The false positive rate was estimated as 3.5% in a 'conditional' setting which means that an abnormality needs to be visible with DBT. As all cancers were visible with DBT,

the combination of mammography with DBT could lead to an increase in cancer detection with a small increase in unjustified referrals (2.0% vs. 3.5%) which is still within the European recommendations. In the Malmö Breast Tomosynthesis Screening Trial [57] one view tomosynthesis is compared to two view mammography. Also the Malmö trial showed an increase in the breast cancer detection rate when using DBT (8.9/1000) compared to mammography (6.3/1000) and an increase in unjustified referrals (2.6/100 to 3.8/100 for mammography and DBT, respectively).

Personalised screening could be implemented in two ways, either by replacing mammography with another modality or by adding additional imaging to mammography. Based on the individual risk for the development of breast cancer and of the risk that the cancer could be missed with mammography, (additional) screening could be offered with MRI or ultrasound. In the US, several studies were conducted to investigated the benefit of an additional MRI or ultrasound examination. The combination of mammography with another screening modality leads to an increase in sensitivity at the cost of more false positive findings [58]. In Japan, a randomised controlled trial was conducted with about 73,000 women [59]. The intervention group got a mammogram and an ultrasonography while the control group underwent mammography only. Also here, more cancers were found when adding an ultrasound examination to mammography and more women were unnecessarily recalled and biopsied. Adding 3D automated breast ultrasound to mammography in women with dense breasts was investigated in a study by Wilczek et al. [60]. More cancers were found when adding 3D ultrasound to mammography, while the recall rate increased slightly and was still within the European recommendations. Additional imaging with MRI for women with extremely dense breasts is currently under investigation within the Dutch population screening in the DENSE trial [61].

In recent years, many states of the US passed legislation that requires the radiologists to inform women about their breast density and possible supplemental imaging options. Melinkow et al. [62] reviewed supplemental screening for breast cancer in women with dense breasts. They found that BI-RADS density assessment is generally consistent across sequential examinations by the same or different readers (at population level), but there was important variability among readings for individual women and that this variability might be reducible with automated assessment. They further state, 'Variability in breast density assignments may lead to unintended consequences. Reclassification from one overall category to another (for example, 'dense' to 'not-dense' or vice versa) may undermine a woman's confidence in the screening process and leave her uncertain about her risk for breast cancer, whereas the opposite reclassification may alarm women unnecessarily or prompt supplemental screening tests of uncertain value.'. Additionally, it is reported that supplemental screening after a negative mammogram leads to additional breast cancer detections, though the positive predictive value is low which means that there are many false positives. In conclusion it is said, that it remains open whether supplemental screening leads to

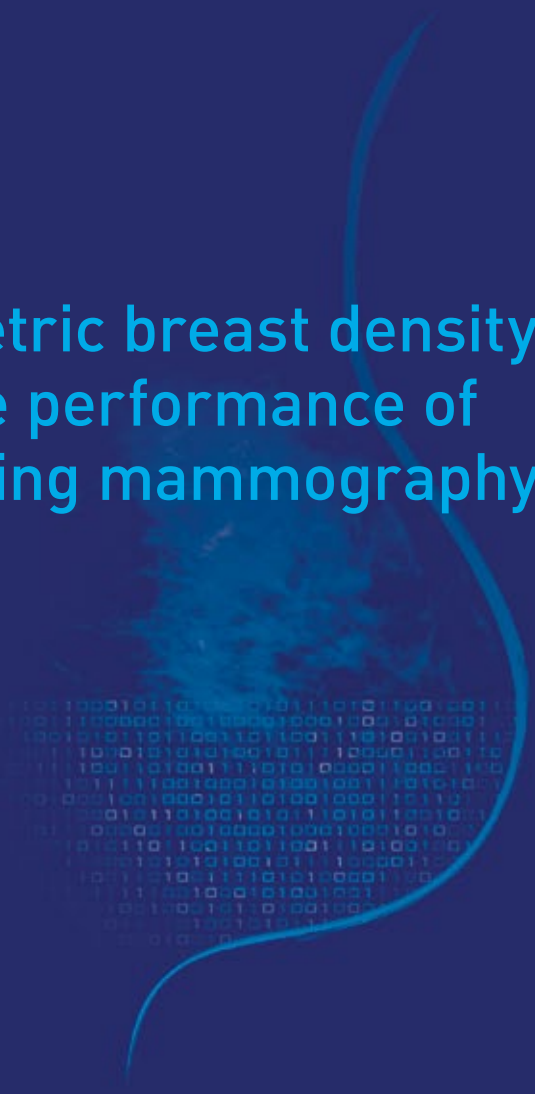
improved clinical outcomes, as only cancer rates are investigated and interval cancer rates and over-diagnosis are not addressed. A need for well-designed, long-term and prospective studies is stated.

1.6 Thesis outline

The outline of the thesis is as follows: First, the Dutch screening program is evaluated considering breast density. Amongst other measurements, the sensitivity and specificity of the program are determined for the four different Volumetric Density Grades (VDG), a scale comparable to the BI-RADS density categories, obtained with the breast density Software Volpara (Volpara Health Technologies, Wellington, New Zealand).
The results of Volpara are also used in Chapter 3. When using breast density for stratification, accurate and reliable measurements are needed over time. The consistency of breast density categories of serial screening mammograms was evaluated. The performance of the software was compared to the performance of human readers using BI-RADS.
Previous research has shown that volumetric breast density is easily underestimated in dense breasts when using an internal calibration to estimate the fibroglandular tissue volume. In Chapter 4, methods to improve breast density estimations in extremely dense breasts are investigated. A pipeline that is suitable for all types of breast densities is proposed.
In dense breasts, cancers are easily masked by dense tissue structures and are therefore not detectable. In Chapter 5, several masking risk estimators are proposed and tested on the ability to distinguish false negative from true negative screening mammograms.
In Chapter 6, the screening program performance is evaluated in relation to the compression pressure applied to the breast during mammogram acquisition. In the analysis, confounding factors like breast density and multiple screening rounds per woman are considered.
This thesis is concluded by summaries and discussions in English, Dutch and German.

2

Volumetric breast density and the performance of screening mammography



Original title: Volumetric breast density affects performance of digital screening mammography.

J.O.P. Wanders, K. Holland, W.B. Veldhuis, R.M. Mann, R.M. Pijnappel, P.H.M. Peeters, C.H. van Gils and N. Karssemeijer

Published in: *Breast Cancer Research and Treatment*, 2017, 162:95-103

Abstract

Purpose: To determine to what extent volumetric mammographic density influences screening performance when using digital mammography.

Methods: We collected a consecutive series of 111,898 digital mammography examinations (2003-2011) from one screening unit of the Dutch biennial screening program (age 50-75 years). Volumetric mammographic density was automatically assessed using Volpara. We determined screening performance measures for four density categories comparable to the American College of Radiology (ACR) breast density categories.

Results: Of all the examinations, 21.6% were categorised as density category 1 ('almost entirely fatty') and 41.5%, 28.9% and 8.0% as category 2 to 4 ('extremely dense'), respectively. We identified 667 screen-detected and 234 interval cancers. Interval cancer rates were 0.7‰, 1.9‰, 2.9‰ and 4.4‰ and false positive rates were 11.2‰, 15.1‰, 18.2‰ and 23.8‰ for categories 1 to 4, respectively (both p-trend$<$0.001). The screening sensitivity, calculated as the proportion of screen-detected among the total of screen-detected and interval tumours, was lower in higher density categories: 85.7%, 77.6%, 69.5% and 61.0% for categories 1 to 4, respectively (p-trend$<$0.001).

Conclusions: Volumetric mammographic density, automatically measured on digital mammograms, impacts screening performance measures along the same patterns as established with ACR breast density categories. Since measuring breast density fully automatically has much higher reproducibility than visual assessment, this automatic method could help with implementing density-based supplemental screening.

2.1 Introduction

Breast density increases breast cancer risk [10, 22]. In addition, sensitivity of screening mammography is lower for women with dense breasts, caused by the masking effect of dense (fibroglandular) breast tissue [26, 27]. This has led to breast density legislation in 28 states of the United States of America (USA) until now, and has fuelled ongoing discussions on the need for supplemental screening for women with dense breasts world-wide [63].
One hoped that screening performance in women with dense breasts would improve when film-screen mammography was replaced by digital mammography. Unfortunately, screening sensitivity was still worse in women with dense compared to non-dense breasts when digital mammography was used [28, 29, 31].
Most large studies looking into the effect of breast density on screening performance used the Breast Imaging-Reporting and Data System (BI-RADS) for breast density assessment, which is assessed by radiologists. However, this method has a moderate inter-observer agreement [40, 42, 64, 65]. With the advent of digital mammography, several fully automatic volumetric density assessment methods have been developed. Volpara is one of these methods, and has shown correlation with BI-RADS density categories and MRI breast density measurements [48–50, 53].
The effect of automatically measured volumetric breast density on screening sensitivity has only been studied once [30]. However, information about the effect of automatically measured volumetric breast density on other screening performance measures like recall rates, false positive rates and positive predictive values (PPV) was not reported in this study. Therefore, the aim of this study was to examine to what extent automatically measured volumetric mammographic density affected screening sensitivity and other screening performance measures in a large Dutch population-based screening program cohort containing a consecutive series of digital screening mammograms and complete information about interval cancers.

2.2 Materials and Methods

2.2.1 Study population

Data were acquired from a breast cancer screening unit (Preventicon screening unit 19, Utrecht, the Netherlands) of the Foundation of Population Screening Mid-West, one of the five screening regions of the Dutch breast cancer screening program. Women participating in this biennial screening program are aged 50 to 75. The program involves mammography only, and all mammograms are read by two certified screening radiologists. In the Dutch screening program, previous screening mammograms are most of the time available for comparison in case of subsequent screens.
In 2003, digital mammography was introduced at the Preventicon screening unit [66–68]. Analogue mammography systems were gradually replaced by digital ones. In July 2007

almost all mammograms at this screening unit were digital [67].
By participating in the Dutch screening program, women consent to their data being used for evaluation and improvement of the screening, unless they have indicated otherwise.

2.2.2 Data collection

We prospectively collected all unprocessed digital mammography examinations that were taken at the Preventicon screening unit between 2003 and 2011, with exception of a 4-month period in 2009 when only processed data were archived. All mammograms were acquired using Lorad Selenia systems (Hologic, Danbury, Conn.). The first screening examination of a woman in the screening program, always included the two standard views, craniocaudal (CC) and mediolateral oblique (MLO). At subsequent screening examinations, MLO was the routinely acquired view and CC was acquired in 57% of the cases by indication (e.g. high breast density, visible abnormality) during the study period. Recall and breast cancer detection information was obtained from the screening registration system. Interval cancers were identified through linkage with the Netherlands Cancer Registry.
Examinations were excluded, when information about recall or final outcome was missing. In addition, examinations for which breast density could not be determined, and interval cancers diagnosed more than 24 months after the last screening mammogram were excluded for analysis.
Tumour information such as maximum diameter, nodal status and ICD-O codes were obtained from the screening registration system. Nodes were classified negative when the sentinel lymph node, or the dissection specimen in case no sentinel lymph node procedure was performed, contained no or only isolated tumour cells. Nodes were considered positive if they contained micrometastases (0.2-2mm) or metastases larger than 2mm.

2.2.3 Volumetric mammographic density assessment

Percentage dense volume (PDV) was automatically assessed from unprocessed mammograms of the left and right breasts, and MLO and CC views using the commercially available Volpara Density software (version 1.5.0, Volpara Solutions, Wellington, New Zealand) [47]. The average PDV per screening examination was determined using the available views of both breasts. Volpara Density Grades (VDGs) were constructed based on this average PDV (VDG 1: $0\% \leq \text{PDV} < 4.5\%$, VDG 2: $4.5\% \leq \text{PDV} < 7.5\%$, VDG 3: $7.5\% \leq \text{PDV} < 15.5\%$, VDG 4: $\text{PDV} \geqslant 15.5\%$). The VDGs are designed to mimic the American College of Radiology BI-RADS breast density categories (4th edition).

2.2.4 Statistical analysis

Examinations were grouped according to VDGs. Within these groups, we determined the following screening performance measures with accompanying 95% confidence intervals (CI) using generalized estimating equations to account for correlation between examina-

tions of the same woman using the 'independence' correlation structure: recall rate, false positive rate, screen-detected breast cancer rate, interval breast cancer rate, total breast cancer rate (all rates are per 1000 screening examinations), sensitivity, specificity and positive predictive value (PPV). For the screening sensitivity, we calculated Wilson's 95% confidence intervals (see Table 2.1 for screening performance definitions). For comparison with American screening programs, we also determined interval cancer rates for the first year after a negative screening mammogram, since the screening interval in the USA is normally 1 year.

We performed several sensitivity analyses: 1) taking only invasive tumours into account (i.e. excluding the examinations leading to a true positive or false negative diagnosis of in situ carcinoma); 2) taking only subsequent screening rounds into account, since performance measures are expected to be different between first and subsequent rounds (in case of subsequent rounds, the prior mammogram could be analogue or digital); 3) using VDGs based on the mean PDV of only the MLO views instead of using all available views.

We tested for linear trends across the four density categories for screening performance measures, the percentage of in situ cancers, and positive lymph nodes with a chi square linear trend test. In addition, we examined whether tumours diagnosed in dense breasts were larger than in non-dense breasts, using the Jonckheere-Terpstra test, as we expected tumour size not to be normally distributed. All statistical tests were two-sided. Statistical analyses were performed in IBM SPSS statistics, version 21 and in R, version 3.2.2 using the 'geese' function from the 'geepack' package.

2.3 Results

In total, 113,956 screening examinations were available. We excluded 50 examinations of which the screening outcome was unknown, 47 interval cancers which were diagnosed more than 24 months after the last screening examination, and 1,961 examinations for which VDG could not be assessed. This resulted in 111,898 examinations belonging to 53,239 women with a median age of 58 years (IQR: 53 - 64 years). Among the examinations, 21.6% were categorised as density category 1 ('almost entirely fatty'), and 41.5%, 28.9% and 8.0% as category 2 to 4 ('extremely dense'), respectively (Table 2.2). In total, 667 screen-detected breast cancers were identified based on a mammogram taken before January 1, 2012, and 234 interval cancers were identified within 24 months after a mammogram taken before January 1, 2012, of which 79.5% and 97.9%, respectively were invasive breast cancers (Table 2.2 and Table 2.3).

2.3.1 Screening performance across volumetric density categories

Table 2.4 shows that total and interval breast cancer rates, recall rates, and false positive rates were higher in higher breast density categories compared to lower density categor-

Interval breast cancers also called false negatives (FN)	Breast cancers diagnosed within 24 months after a screening examination that did not lead to recall (negative mammogram), and before the next scheduled screening examination.
Screen-detected breast cancers also called true positives (TP)	Breast cancers diagnosed after a recalled screening examination (positive mammogram).
False positives (FP)	Screening examinations that led to a recall (positive mammogram), but not to a breast cancer diagnosis within 24 months after the examination, or before the next scheduled screening examination.
True negatives (TN)	Screening examinations that did not lead to recall (negative mammogram) and no breast cancer was diagnosed within 24 months after the examination, or before the next scheduled screening examination.
Sensitivity of screening	The number of screen-detected breast cancers divided by the total number of screen-detected plus interval breast cancers ((TP/(TP+FN)).
Specificity of screening	Number of screening examinations that did not lead to recall (negative mammogram) and no breast cancer diagnosis within 24 months, or before the next scheduled screening examination divided by the total number of examinations without breast cancer diagnosis within 24 months, or before the next scheduled screening examination ((TN/(TN+FP)).
Positive predictive value (PPV)	The number of screen-detected breast cancers divided by the total number of examinations that led to recall ((TP/(TP+FP)).

Table 2.1: Definitions of screening performance measures.

ies, all with a significant linear trend (p-trend<0.001). Screen-detected breast cancer rates were found to be lowest in the lowest breast density category (4.0 per 1000 examinations (‰)) and more comparable across the three highest breast density categories: 6.4‰, 6.6‰ and 6.8‰, respectively (p-trend<0.001). The screening sensitivity was significantly lower (p-trend<0.001) in higher breast density categories: 85.7%, 77.6%, 69.5%, 61.0% in VDG categories 1 to 4, respectively. No significant linear trend was found for PPV (p-trend=0.12) (Table 2.4).

Overall trends for interval cancer rates, recall rates and false positive rates, screening sensitivity and specificity were similar when either invasive cancers alone or both invasive cancer and in situ cancers, were taken into account. However, when restricting the analyses to invasive cancers only, the screening sensitivity in VDG 4 decreased most notably compared to the screening sensitivity when both in situ and invasive breast cancers were taken into account. When only subsequent screening rounds were taken into account, the overall trends were again similar to the analyses based on both first and subsequent screening examinations (Table 2.4).

The results of the sensitivity analysis where PDV was based on MLO views only did not differ from those based on all available views (data not shown).

In VDG category 1, 25% of the interval breast cancers were diagnosed in the first year after screening examination; in VDG categories 2 and 3 this was 41% and in VDG category 4 67%. This resulted in interval cancer rates in the first year after a screening examination of 0.2‰, 0.8‰, 1.2‰, and 2.9‰ (p-trend<0.001) in VDG categories 1, 2, 3, and 4, respectively.

2.3.2 Tumour characteristics across volumetric density categories

Of all tumours 74.0% were screen-detected and 26.0% were interval cancers. 15.7% of all tumours were in situ and 84.3% were invasive tumours. 89.4% of the in situ tumours showed microcalcifications on the last screening mammogram. For screen-detected tumours, the highest proportion of in situ tumours was found in the highest density category (in VDG 4, 32.8% of the screen-detected tumours were in situ tumours) and the lowest proportion in density category 2 (in VDG 2, 15.8% of the screen-detected tumours were in situ tumours). A significant linear trend was observed for the proportion of invasive tumours over breast density categories among screen-detected tumours (p-trend=0.03).

About 80% of the screen-detected and slightly over 50% of the interval invasive breast cancers were smaller than 20mm (pT1 status) at diagnosis. No linear trend was found for screen-detected tumour size across the four density categories (p-trend$_{SD}$=0.10) (Table 2.3). Lymph nodes were positive in 29.3% of the screen-detected cancers and 36.8% of the interval cancers. For lymph node status, no linear trend was found across the four breast density categories for screen-detected breast cancers (p-trend$_{SD}$=0.08) (Table 2.3).

	Total	VDG 1		VDG 2		VDG 3		VDG 4	
All screening rounds									
Screening examinations	111,898	24,210	(21.6%)	46,426	(41.5%)	32,330	(28.9%)	8,932	(8.0%)
Screen-detected cancers	667	96	(14.4%)	298	(44.7%)	212	(31.8%)	61	(9.1%)
Interval cancers	234	16	(6.8%)	86	(36.8%)	93	(39.7%)	39	(16.7%)
False positives	1,774	271	(15.3%)	700	(39.5%)	590	(33.3%)	213	(12.0%)
True negatives	109,223	23,827	(21.8%)	45,342	(41.5%)	31,435	(28.8%)	8,619	(7.9%)
Only invasive tumours									
Screening examinations	111,754	24,188	(21.6%)	46,375	(41.5%)	32,279	(28.9%)	8,912	(8.0%)
Screen-detected cancers	529	75	(14.2%)	250	(47.3%)	163	(30.8%)	41	(7.8%)
Interval cancers	228	15	(6.6%)	83	(36.4%)	91	(39.9%)	39	(17.1%)
False positives	1,774	271	(15.3%)	700	(39.5%)	590	(33.3%)	213	(12.0%)
True negatives	109,223	23,827	(218%)	45,342	(41.5%)	31,435	(28.8%)	8,619	(7.9%)
Only subsequent screening rounds									
Screening examinations	94,665	22,146	(23.4%)	40,664	(43.0%)	25,777	(27.2%)	6,078	(6.4%)
Screen-detected cancers	521	86	(16.5%)	249	(47.8%)	152	(29.2%)	34	(6.5%)
Interval cancers	203	16	(7.9%)	81	(39.9%)	80	(39.4%)	26	(12.8%)
False positives	1,170	214	(18.3%)	491	(42.0%)	366	(31.3%)	99	(8.5%)
True negatives	92,771	21,830	(23.5%)	39,843	(42.9%)	25,179	(27.1%)	5,919	(6.4%)

Table 2.2: Numbers in total and within Volpara Density Grade (VDG) categories (based on the available views). Within brackets is the percentage of each VDG category as compared to the total.

		Total		VDG 1		VDG 2		VDG 3		VDG 4		p-trend
Proportion invasive tumours[a]												
Total (N=898)	invasive	757	(84.3%)	90	(80.4%)	333	(87.2%)	254	(83.6%)	80	(80.0%)	0.49
Screen-detected cancers (N=665)	invasive	529	(79.5%)	75	(78.1%)	250	(84.2%)	163	(77.3%)	41	(67.2%)	0.03
Interval cancers (N=233)	invasive	228	(97.9%)	15	(93.8%)	83	(97.6%)	91	(97.8%)	39	(100.0%)	0.20
pT (only invasive tumours)[b]												
Total (N=700)	T1	503	(71.9%)	70	(81.4%)	231	(73.6%)	153	(66.8%)	49	(69.0%)	0.02[c]
	T2	171	(24.4%)	15	(17.4%)	74	(23.6%)	65	(28.4%)	17	(23.9%)	
	T3 & T4	26	(3.7%)	1	(1.2%)	9	(2.9%)	11	(4.8%)	5	(7.0%)	
Screen-detected cancers (N=511)	T1	404	(79.1%)	63	(85.1%)	195	(79.6%)	116	(75.8%)	30	(76.9%)	0.14[c]
	T2	97	(19.0%)	11	(14.9%)	46	(18.8%)	33	(21.6%)	7	(17.9%)	
	T3 & T4	10	(2.0%)	0	(0.0%)	4	(1.6%)	4	(2.6%)	2	(5.1%)	
Interval cancers (N=189)	T1	99	(52.4%)	7	(58.3%)	36	(52.2%)	37	(48.7%)	19	(59.4%)	0.87[c]
	T2	74	(39.2%)	4	(33.3%)	28	(40.6%)	32	(42.1%)	10	(31.3%)	
	T3 & T4	16	(8.5%)	1	(8.3%)	5	(7.2%)	7	(9.2%)	3	(9.4%)	
Lymph node status (only invasive tumours)[d]												
Total (N=741)	positive	234	(31.6%)	18	(20.2%)	105	(32.3%)	87	(35.2%)	24	(30.0%)	0.12
Screen-detected cancers (N=518)	positive	152	(29.3%)	13	(17.6%)	75	(30.7%)	51	(32.1%)	13	(31.7%)	0.08
Interval cancers (N=223)	positive	82	(36.8%)	5	(33.3%)	30	(37.0%)	36	(40.9%)	11	(28.2%)	0.68
Tumour diameter (only invasive tumours)[e]												
Total (N=691)	Median (mm) (IQR)	15	(10-22)	12	(8-18)	15	(10-21)	17	(11-25)	14	(10-22)	0.01
Screen-detected cancers (N=500)	Median (mm) (IQR)	13	(9-19)	11	(8-17)	13	(10-19)	14	(10-20)	12	(8-19)	0.10
Interval cancers (N=191)	Median (mm) (IQR)	20	(14-30)	20	(13-33)	19	(16-30)	21	(16-31)	16	(12-25)	0.34

a) Information on invasiveness is missing for 3 tumours (2 screen-detected & 1 interval tumours); b) Information on pT status is missing for 57 tumours (18 screen-detected & 39 interval tumours); c) p-trend determined for T1 versus T2, T3, and T4; d) Information on lymph node status is missing for 16 tumours (11 screen-detected & 5 interval tumours); e) Information on tumours diameter is missing for 66 tumours (29 screen-detected & 37 interval tumours)

Table 2.3: Tumour characteristics in total and within Volpara Density Grade (VDG) categories (based on the available views).

	Total		VDG 1		VDG 2		VDG 3		VDG 4		p-trend
All screening rounds											
Recalls /1000	21.8	(20.9-22.7)	15.2	(13.7-16.8)	21.7	(20.2-22.9)	24.8	(23.1-26.6)	30.7	(27.2-34.5)	<0.001
False positives /1000	15.9	(15.1-16.6)	11.2	(9.9-12.6)	15.1	(14.0-16.2)	18.2	(16.8-19.8)	23.8	(20.8-27.3)	<0.001
Screen-detected cancers /1000	6.0	(5.5-6.4)	4.0	(3.2-4.8)	6.4	(5.7-7.2)	6.6	(5.7-7.5)	6.8	(5.3-8.8)	<0.001
Interval cancers /1000	2.1	(1.9-2.4)	0.7	(0.4-1.1)	1.9	(1.5-2.3)	2.9	(2.3-3.5)	4.4	(3.2-6.0)	<0.001
Breast cancers /1000	8.1	(7.6-8.7)	4.6	(3.8-5.6)	8.3	(7.5-9.1)	9.4	(8.4-10.5)	11.2	(9.2-13.6)	<0.001
Sensitivity of screening (%)	74.0	(71.1-76.7)	85.7	(78.1-91.0)	77.6	(73.2-81.5)	69.5	(64.1-74.4)	61.0	(51.2-70.0)	<0.001
Specificity (%)	98.4	(98.3-98.5)	98.9	(98.7-99.0)	98.5	(98.4-98.6)	98.2	(98.0-98.3)	97.6	(97.2-97.9)	<0.001
Positive predictive value(%)	27.3	(25.6-29.1)	26.2	(21.9-30.9)	29.9	(27.1-32.8)	26.4	(23.5-29.6)	22.3	(17.7-27.6)	0.12
Only invasive tumours											
Recalls /1000	20.6	(19.8-21.4)	14.3	(12.9-15.9)	20.5	(19.2-21.8)	23.3	(21.7-25.1)	28.5	(25.2-32.3)	<0.001
False positives /1000	15.9	(15.1-16.6)	11.2	(9.9-12.6)	15.1	(14.0-16.3)	18.3	(16.9-19.8)	23.9	(20.9-27.4)	<0.001
Screen-detected cancers /1000	4.7	(4.3-5.1)	3.1	(2.5-3.9)	5.4	(4.8-6.1)	5.0	(4.3-5.9)	4.6	(3.4-6.2)	0.02
Interval cancers /1000	2.1	(1.9-2.4)	0.6	(0.4-1.0)	1.8	(1.4-2.2)	2.8	(2.3-3.5)	4.4	(3.2 ;6.0)	<0.001
Breast cancers /1000	6.9	(6.4-7.3)	3.7	(3.0-4.6)	7.2	(6.5-8.0)	7.9	(7.0-8.9)	9.0	(7.2-11.1)	<0.001
Sensitivity of screening (%)	69.1	(66.5-73.0)	83.3	(74.3-89.6)	74.4	(70.2-79.4)	62.9	(58.1-69.8)	50.6	(40.5-61.9)	<0.001
Specificity (%)	98.4	(98.3-98.5)	98.9	(98.7-99.0)	98.5	(98.4-98.6)	98.2	(98.0-98.3)	97.6	(97.2-97.9)	<0.001
Positive predictive value(%)	23.0	(21.3-24.7)	21.7	(17.6-26.3)	26.3	(23.6-29.2)	21.6	(18.9-24.7)	16.1	(12.1-21.2)	0.02

Table continues on the next page.

	Total		VDG 1		VDG 2		VDG 3		VDG 4		p-trend
Only subsequent screening rounds											
Recalls /1000	17.9	(17.0-18.7)	13.5	(12.1-15.2)	18.2	(16.9-19.5)	20.1	(18.4-21.9)	21.9	(18.5-25.9)	<0.001
False positives /1000	12.4	(11.7-13.1)	9.7	(8.4-11.0)	12.1	(11.1-13.2)	14.2	(12.8-15.7)	16.3	(13.3-19.9)	<0.001
Screen-detected cancers /1000	5.5	(5.0-6.0)	3.9	(3.1-4.8)	6.1	(5.4-6.9)	5.9	(5.0-6.9)	5.6	(4.0-7.8)	0.02
Interval cancers /1000	2.2	(1.9-2.5)	0.7	(0.4-1.2)	2.0	(1.6-2.5)	3.1	(2.5-3.9)	4.3	(2.9-6.3)	<0.001
Breast cancers /1000	7.7	(7.2-8.3)	4.6	(3.8-5.6)	8.1	(7.3-9.0)	9.0	(7.9; 10.2)	9.9	(7.7-12.7)	<0.001
Sensitivity of screening (%)	71.3	(68.6-75.1)	84.3	(76.0-90.1)	74.8	(70.1-79.8)	64.4	(59.2-71.3)	56.7	(44.1-68.4)	<0.001
Specificity (%)	98.8	(98.7-98.8)	99.0	(98.9-99.2)	98.8	(98.7-98.9)	98.6	(98.4; 98.7)	98.4	(98.0-98.7)	<0.001
Positive predictive value (%)	30.8	(28.6-33.0)	28.7	(23.8-34.0)	33.6	(30.3-37.1)	29.3	(25.6-33.4)	25.6	(18.8-33.7)	0.35

Table 2.4: Screening performance measures (and 95% CI) in total and within Volpara Density Grade (VDG) categories (based on the available views).

2.4 Discussion

We found that the sensitivity of a digital mammography screening program was significantly lower in women with high volumetric breast density than in women with low volumetric breast density (61.0% and 85.7%, respectively, (p-trend<0.001)). This is despite the higher recall rates in women with high compared to low breast density (30.7‰ and 15.2‰, respectively) (p-trend<0.001).

A study of Destounis et al. [30], which was recently published, also studied the screening sensitivity in four automatically determined volumetric breast density categories. They found screening sensitivities of 95%, 89%, 83% and 65% in density categories 1 to 4 respectively. Additionally, they determined the mammographic screening sensitivity across the visual BI-RADS categories and found sensitivities of 82% in the lowest and 66% in the highest breast density category.

Four other studies where breast density was visually assessed on digital screening mammograms, also found a negative influence of breast density on screening sensitivity [28, 29, 31, 32], a fifth study [69] did not find this result. A Canadian study [31] showed a lower screening sensitivity for women with 75% or higher breast density (74.2% (95% CI: 67.2-80.4)) compared to women with less than 75% breast density (80.2% (95% CI:78.4-81.9)) when using direct radiography in a biennial screening program, where women who are considered to be at increased risk were screened annually. In the American Digital Mammographic Imaging Screening Trial (DMIST), the screening sensitivity was determined for women with dense and non-dense breasts for several subgroups. Sensitivity seemed higher for all non-dense compared to dense subgroup comparisons, with exception of postmenopausal women aged 50 to 64 years [28]. In a study using data from the Breast Cancer Surveillance Consortium (BCSC), Kerlikowske et al. [29] found that in an annual digital mammography screening program sensitivity was also significantly lower in the higher BI-RADS breast density categories than in the lower BI-RADS categories for women aged 50 to 74 years. However, in another paper by Kerlikowske et al. [69], also using BCSC data, no significant differences in screening sensitivity between breast density categories was found, when digital mammography was used. Finally, in a recently published study of Weigel et al. [32], where data of the German biennial screening program was used, screening sensitivity was found to be lower in the higher as compared to the lower breast density categories. In that study, screening sensitivities of 100% and 50% were found for the lowest and the highest density category, respectively.

Although the results in the above studies are not completely consistent, the majority of them showed that screening performance is still negatively influenced by breast density when digital mammography is used instead of film screen mammography. This is also found in the current study.

Four out of six above-mentioned studies were conducted in the USA [28–30, 69]. The only European study determining the influence of breast density on digital mammography

screening performance was the recently published study of Weigel et al. [32]. However, our study is the first to determine the effect of automatically assessed volumetric mammographic density on digital mammography screening performance in a European population-based screening setting. There are three notable differences between European and American screening programs: 1) recall rates are below 5-7% in Europe and around 8-10% in the USA [69–75]; 2) double reading, which is also used in this study, is common in European screening programs, but not in the USA [76]; 3) the screening interval is different. Biennial screening is common in European countries, while in the USA women are mostly screened yearly [76].

When looking at the interval cancers diagnosed within the first year after a negative screening mammogram, we found that in the lower density categories only a small part of the interval cancers were found in the first year after a negative screening examination, and most were found in the second year, whereas in women with extremely dense breasts, this was the other way around. Although a one year screening interval instead of a 2 year screening interval would probably result in a higher program sensitivity in all density groups, this will happen to a larger extent in the women with fatty breasts than in those with extremely dense breasts, resulting in larger differences in screening sensitivity across density categories.

When only invasive cancers instead of both invasive and in situ cancers were taken into account, the screening sensitivity decreased most notably in VDG 4. This indicates that the detection of invasive breast cancers in digital mammography screening is hampered to a larger extent than the detection of in situ breast cancers (Table 2.4). A possible explanation for this is that the visibility of microcalcifications, that often are the hallmark of ductal carcinoma in situ (DCIS) on mammography [68], is not hampered as much in dense tissue as the visibility of invasive breast cancers. 89.4% of the DCIS in our study was accompanied by microcalcifications.

False positive rates were found to be higher in women with dense breasts compared to women with non-dense breasts. Similar trends were found in two American studies using BCSC data [29, 77].

When looking at the tumour characteristics of screen-detected breast cancers, we observe a significant linear trend for the proportion of invasive tumours over breast density categories (p-trend=0.03). In addition, the size of screen-detected cancers, and the proportion of positive lymph node status among screen-detected cancers seem to be larger in denser breasts. However, no significant linear trend was found for screen-detected tumour size and positive lymph node status proportion across the four density categories (p-trend size=0.10 & p-trend lymph node status=0.08).

It should be noted that the four density categories (VDGs) used in this study are comparable to the 4th edition BI-RADS density categories. Although in 2013 the 5th BI-RADS density edition was introduced, we here still used the VDG categories comparable to the

4th edition, to enable better comparison with previous studies.

A limitation of this study is that during the study period the MLO view was the standardly acquired view for the subsequent screening rounds and CC views were only taken in addition to MLO during the first screening round or by indication during subsequent rounds. As a result, breast density was determined based on only MLO views for some examinations and on both MLO and CC views for other examinations in our main analysis. Volpara's PDV measured on CC views tends to be somewhat higher than on MLO views [24]. As CC views are more often performed among women with dense breasts and women with a suspicious region on their MLO view, breast density might be somewhat artificially elevated for these women. Our sensitivity analysis using VDG categories based on PDV from the MLO views only did not lead to different conclusions. Screening sensitivity is presumably higher when both MLO and CC views are available compared to MLO views only. Therefore, standardly taking both MLO and CC views would lead to higher sensitivity, particularly in women with fatty breasts as they are the ones who most often receive MLO views only. This would lead to larger differences in screening performance across breast density categories.

Strengths of this study are the large sample size and the fact that the digital mammograms were acquired in routine screening. In addition, we used a fully automatic method to determine PDV, which was possible because unprocessed image data were archived. In several studies, this automatic method (Volpara) showed to be correlated with BI-RADS breast density and to give comparable breast cancer risk estimations as with BI-RADS breast density [48–50]. In addition, it has been validated against MRI [53]. Volpara gives objective and reproducible density measurements, representing the amount of dense tissue rather than the size of the dense tissue projection as measured by area-based methods.

In summary, in a large screening population, where digital mammography was used for screening and a fully automatic method (Volpara) was used to determine PDV, breast density was found to significantly hamper the detection of breast tumours. This is shown by a lower screening sensitivity in women with dense compared to those with non-dense breasts, which existed despite a higher recall rate for women with dense breasts. These findings are in line with results of most studies using visually assessed BI-RADS density on digital mammograms. Since measuring breast density fully automatically has higher reproducibility than visual assessment, this automatic method could help with facilitating a more tailored screening, such as supplemental screening for women with dense breasts.

3

Consistency of breast density categories in serial screening mammograms

Original title: Consistency of breast density categories in serial screening mammograms: A comparison between automated and human assessment.

K. Holland, J. van Zelst, G.J. den Heeten, M. Imhof-Tas, R.M. Mann, C.H. van Gils and N. Karssemeijer

Published in: *The Breast,* 2016, 29:49-54

Abstract

Reliable breast density measurement is needed to personalise screening by using density as a risk factor and offering supplemental screening to women with dense breasts. We investigated the categorisation of pairs of subsequent screening mammograms into density classes by human readers and by an automated system.

With software (VDG) and by four readers, including three specialised breast radiologists, 1000 mammograms belonging to 500 pairs of subsequent screening exams were categorised into either two or four density classes. We calculated percent agreement and the percentage of women that changed from dense to non-dense and vice versa. Inter-exam agreement (IEA) was calculated with kappa statistics. Results were computed for each reader individually and for the case that each mammogram was classified by one of the four readers by random assignment (group reading).

Higher percent agreement was found with VDG (90.4%, 95% CI 87.9-92.9%) than with readers (86.2-89.2%), while less plausible changes from non-dense to dense occur less often with VDG (2.8%, 95% CI 1.4-4.2%) than with group reading (4.2%, 95%CI 2.4-6.0%). We found an IEA of 0.68-0.77 for the readers using two classes and an IEA of 0.76-0.82 using four classes. IEA is significantly higher with VDG compared to group reading.

The categorisation of serial mammograms in density classes is more consistent with automated software than with a mixed group of human readers. When using breast density to personalise screening protocols, assessment with software may be preferred over assessment by radiologists.

3.1 Introduction

The association between breast density and breast cancer risk is well established. Several studies show that the risk of developing breast cancer is two to six times higher for women with dense breasts than for women in the lowest density category [10, 22–24]. Though studies suggest that with the introduction of digital mammography differences in sensitivity across density categories disappear [69], sensitivity of mammography is still impaired by density, because dense tissue can mask cancers [29, 31]. Therefore, personalised breast cancer screening protocols involving ultrasound and MRI are developed taking into account breast density [34].

The most common breast density reporting method is the Breast Imaging Reporting and Data System (BI-RADS) [37] which uses four categories. Studies have shown a considerable inter- and intra-reader variability when using BI-RADS [39–42].

To overcome these variabilities, semi and fully automatic methods were developed to quantify breast density. A first approach was the area based method Cumulus [55]. With Cumulus, the radiologist has to set a threshold to distinguish fibroglandular tissue from fatty tissue. Subsequently, the proportion fibroglandular tissue is calculated with respect to the breast area. BI-RADS and Cumulus are limited by the fact that they are based on the two-dimensional projection of fibroglandular tissue. This projection varies with the projection angle and threshold settings and ignores the three-dimensional anatomical breast structure. To overcome these limitations, quantitative image analysis methods were developed based on imaging physics [45–47, 78, 79]. These methods take the thickness of the compressed breast and imaging parameters into account to measure the absolute (cubic centimetres) and relative (percentage of the breast volume) amount of fibroglandular tissue.

Development of automated breast density assessment methods is an important step towards the introduction of personalised screening protocols adjusted to the need of individual women. This includes supplemental screening to mammography or the replacement of mammography with MRI or ultrasound. To be accepted in practice, it is important to have a consistent, objective and reproducible measurement of breast density to stratify women unambiguously in non-dense and dense categories. With a poor density measurement clinicians and women may lose confidence in the stratification process. Therefore, the temporal aspect of density measurements is very important, as it may be hard to explain why supplemental screening is offered in an irregular pattern. This is acknowledged in a recent review paper, where concerns are raised that radiologists‘ variability of BI-RADS density assessments over time may lead to inconsistent information in mandated communications about elevated breast cancer risk and supplemental screening [62].

Changes in density classes over time can be caused by changes in hormonal status or a change of BMI. It is known that density usually decreases gradually with lifetime [25], so a change to a lower category is expected for some women. The reproducibility of automated volumetric breast density measurements was studied with repeated exams [80, 81].

In these studies no significant differences in density measurements were observed. Furthermore, several studies found good correlations between automated and human density assessment [49, 50, 64].
The purpose of this study is to investigate the consistency of density classifications in serial screening mammograms with fully automated volumetric density measurements and to compare these results to classifications of human readers, operating individually or as a group with mammograms distributed randomly over the readers. The latter does better reflect screening practice as serial mammograms are usually not read by the same radiologist.

3.2 Materials and Methods

3.2.1 Materials

Digital Mammograms from the Dutch breast cancer screening program were used which were acquired in a population of 56,000 women between 2003 and 2012. In this program, women aged 50-75 receive a biennial invitation for breast cancer screening. All mammograms were recorded with Lorad Selenia systems (Hologic, Bedford, USA). Consecutive exam pairs were selected in which we call the oldest exam the prior and the more recent exam the current. All mammograms were processed with Volpara (v1.5.0, Volpara Health Technologies, Wellington, New Zealand) to obtain breast density scores. For this purpose we used the 'for processing' (raw) data. In total, there were 67,260 pairs and for 64,308 pairs density computation was successful. Missing values were due to breast implants (1.3%) and software failures (3.1%). For this study, we randomly selected 500 women, where for every woman one pair with a prior and current exam was selected at random.
The average screening interval in the 500 pairs was 30 months and is more than 24 months, because sometimes women skip screening. A screening interval of 26 months was measured most frequently, which corresponds to the median screening interval. The mean age was 58.8$\pm$6.7 years at the prior screening.
Not all mammograms in our study had four views, because until recently four-view mammography was not standard in the Dutch screening program. Instead, four views were taken in the first screening round and in subsequent rounds only mediallateral oblique (MLO) images were acquired unless there was an indication for additional cranialcaudal (CC) images, like high breast density or a possible abnormality judged by the radiographer. Of the 1000 exams used, 415 exams had MLO views only, while 585 exams had MLO and CC views. For 473 exams, only 'for processing' images were available. To enable density assessment by the radiologists these exams were converted to 'for presentation' format using dedicated software (called 'in-house' processing through this paper). It was verified that the presentation quality of these images was appropriate for density assessment.

3.2.2 Experimental design

For all images, volumetric percent density was calculated with Volpara by dividing the fibroglandular tissue volume by the breast volume. Volpara uses quantitative image analysis algorithms based on physical models [45–47]. We averaged all available percent density estimations of an exam. Using the averaged percent density estimate, we categorised all exams using the Volpara density grade (VDG) [82], which is a four point scale matched to the BI-RADS categories. Additionally, we categorised studies with a VDG of one or two as non-dense while we labelled studies with VDG of three or four as dense.

Three radiologists (R1, R2 and R3) with more than eight years of experience in breast imaging and a PhD student (R4) with a medical degree and two years of experience with breast imaging assigned BI-RADS scores (4th edition) individually to each exam. The radiologists were familiar with the density categories, as these are routinely assessed in clinical practice. We categorised studies with a score of one or two as non-dense while studies with a score of three or four were categorised as dense.

Each reader performed the BI-RADS scoring in two reading sessions with at least one week between the sessions. In each of the sessions, 500 exams were scored, including either the prior or the current mammogram of a pair. Each of these sessions contained 250 prior and 250 current exams. In screening practice current and prior mammograms are often read by different radiologists. Therefore, we also constructed a 'group reading', abbreviated with RG, by assigning the score of a randomly chosen reader to each exam.

We organised two additional reading sessions to study the intra-reader variability and the effect of processing. An overview of these third and fourth reading sessions is shown in Figure 3.1. We selected 50 exams with 'in-house' processing and 50 exams with the original processing from the 1000 exams read in the first two sessions. These 100 exams were chosen randomly, though with the same VDG distributions as the 1000 exams. The selected exams were equally distributed over the two sets presented in the third and fourth reading session. Additionally, the 50 exams that were read in the first sessions with original processing were duplicated in the series to be read, but now with in-house processing. The exams were distributed over the reading sessions as well, where we took care that the original and the in-house processing of an exam were not presented in the same session. Thus, 75 exams were scored in each session. This experimental setting allows us to assess the intra-reader agreement by comparing the density scores given in the two additional sessions to the density scores given in first two reading sessions. Secondly, we can compare the density score given based on the reading with in-house processing to the density score given based on images presented with original processing (Figure 3.2).

3.2.3 Statistical methods

The percentage of women categorised in the same class for the prior and current exam and the percentage of women that change from the non-dense to the dense category and vice

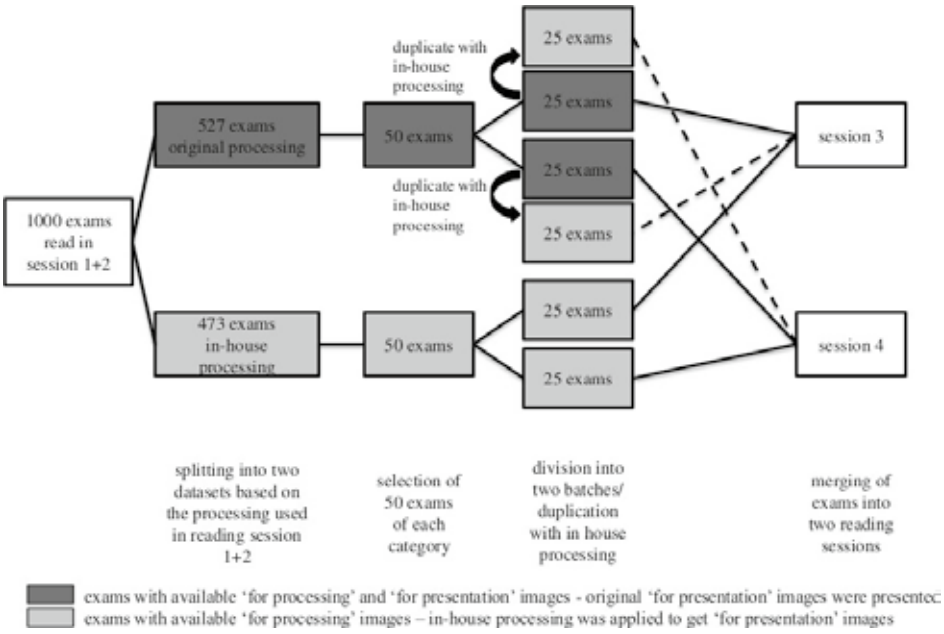


Figure 3.1: An overview of the procedure to study the intra-reader variability and the effect of processing.

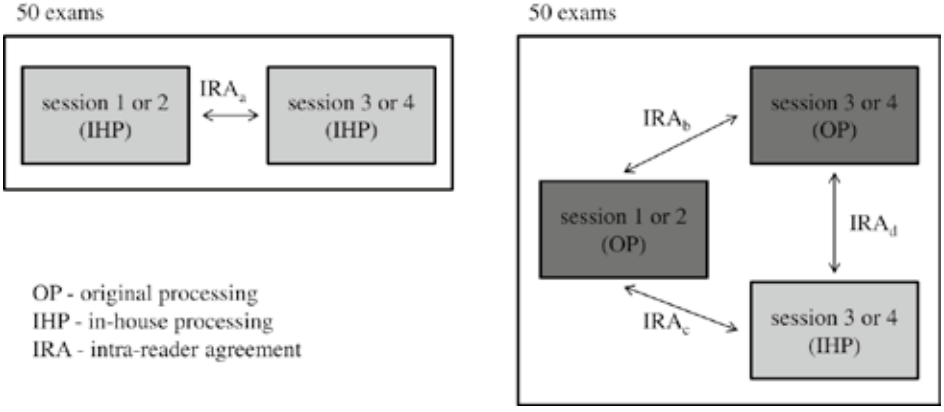


Figure 3.2: An overview about the possible intra-reader agreement calculations. For the 50 images that were read in session 1 and 2 with original processing, three different intra-reader agreements calculations are possible. For the image that were initially read with in-house processing, only one intra-reader agreement was determined.

versa were calculated using the results of the first two reading sessions. We used bootstrapping to calculate the 95% confidence interval (CI) of these percentages and to determine if there are significant differences between VDG and results of each reader.
The inter-exam agreement was calculated with Cohen's weighted kappa, using either two or four density classes. Moreover, we determined the average kappa value of the readers. To compare the kappa value of the group reader and VDG we used bootstrapping [83]. We also calculated the agreement for a subset of cases in which women who skipped a screening round were excluded. We selected this subset by including only cases for which the screening interval between the exams was within 20-28 months.
For comparison with other studies in the literature, intraclass correlation coefficients (ICCs) were calculated to examine the inter-exam agreement of the four classes categorisation.
Additionally, the inter-reader agreement was determined using kappa statistics and the scores assigned to the 1000 exams of the first two reading sessions.
Intra-reader agreements for different reading conditions, depending on processing type, were determined with Cohen's weighted kappa, taking into account two or four density classes. Figure 3.2 gives an overview of the comparisons we made. First, for the images that were presented with in-house processing the intra-reader agreement (IRA_a) between scores in the first two sessions and those in the third and fourth sessions was calculated. Second, the same was done for the 50 exams that were presented with original processing in these sessions (IRA_b). Third, we determined intra-reader agreement for the condition that processing types were different, by comparing the scores of the 50 cases with original processing in the first two sessions that were presented again, but with in-house processing, in the additional session (IRA_c). Fourth, we determined intra-reader agreement for the mixed processing condition using the exams presented twice, once with in-house and once with original processing, in the additional reading sessions (IRA_d). All four kappa values were compared to each other including correction for multiple testing (Bonferroni).
We used the kappa definition of Landis and Koch [84], with the following interpretation: values <0 poor agreement; 0.00-0.20 slight agreement; 0.21-0.40 fair agreement; 0.41-0.60 moderate agreement; 0.61-0.80 substantial agreement; and 0.81-1.00 represent almost perfect agreement. The statistical analysis was performed using the statistical software package R (v3.1.1, R Foundation for Statistical Computing, Vienna, Austria). Starting point of our analysis was a script of Vanbelle [85].

3.3 Results

3.3.1 Inter-exam agreement

Breast density assessments are compared by looking at changes of the scores between prior and current exams (500 pairs). In Table 3.1 we show changes using the two category classification of dense and non-dense mammograms and the percentage of cases without change

including 95% confidence intervals. VDG has the highest percentage of cases without change, 90.4% (CI 87.9-92.9%) compared to 86.2-89.2% for the readers, which is significantly higher than reader 3 and the group reading. The percentage of cases with a change from non-dense to dense was 2.8% (CI 1.4-4.2%) using VDG, which was lower than the 4.2% (CI 2.4-6.0%) obtained with group reading. Reader 4 had the same result as VDG and only reader 1 had a lower score. The percentage of women in which a category change from dense to non-dense was observed ranged from 6.6% to 10.4%.

	non-dense prior and dense current			dense prior and non-dense current			no category change		
VDG	2.8	(1.4-4.2)	-	6.8	(4.6-9.0)	-	90.4	(87.9-92.9)	-
R1	1.4	(0.4-2.4)	0.077	9.7	(7.0-12.4)	0.067	89.0	(86.3-91.7)	0.379
R2	5.2	(3.2-7.2)	0.043	6.6	(4.4-8.8)	0.860	88.2	(85.5-90.9)	0.204
R3	3.4	(1.8-5.0)	0.589	10.4	(7.7-13.1)	0.030	86.2	(83.1-89.3)	0.029
R4	2.8	(1.4-4.2)	0.986	8.0	(5.6-10.4)	0.396	89.2	(86.5-91.9)	0.456
RG	4.2	(2.4-6.0)	0.222	9.1	(6.6-11.6)	0.157	86.8	(83.9-89.7)	0.038

Table 3.1: Mean percentage pairs with a category change from non-dense to dense, dense to non-dense and mean percentage pairs that got twice the same density score (using two categories) of the 500 pairs. The 95% CI is given in brackets and the p-value for the comparison between VDG and the individual readers (R1-R4), and between VDG and the group reading (RG) given as well.

Consistency of breast density in serial mammograms was also evaluated by calculating kappa values. The agreement is given in Table 3.2 column (a). The agreement was substantial for the readers with values ranging from 0.76-0.82 and 0.68-0.77 using four and two classes, respectively. Using VDG we obtained a kappa of 0.85 (CI 0.82-0.87) and 0.80 (CI 0.74-0.85) for four and two classes, respectively. The kappa value is significantly higher with VDG than with group reading ($p<0.001$ and $p=0.010$ for four and two classes).
Table 3.2 also shows the agreement for a subset of 373 women (screening interval 20-28 months). The agreement is slightly higher on average, because women who skipped a screening were excluded in this subset. These women are expected to have a larger decrease in breast density caused by normal involution.
The ICC (95% CI) of the scores for the prior and current exams was 0.91 (0.89-0.92), 0.79 (0.75-0.82), 0.77 (0.73-0.81), 0.76 (0.72-0.79), 0.82 (0.79-0.84), and 0.75 (0.71-0.78) for VDG, R1, R2, R3, R4 and RG, respectively. The contingency tables of VDG and the reader scores of prior and current exams are shown in Table 3.4.

3.3.2 Inter-reader agreement

With the scores of the 1000 exams presented in the two sessions the inter-reader agreement was calculated. We found a substantial to almost perfect agreement, with kappa values

between 0.78 and 0.83 using four categories. The agreement for two categories is between 0.73 and 0.78. The agreement between the readers and VDG is lower. We observed kappa values between 0.73 and 0.78, and 0.63 and 0.71 using four and two categories, respectively. In most of the pairs with a disagreement between VDG and the reader, a higher score was given by the software than by the reader. The individual inter-reader agreements are given in Table 3.3.

A) four categories			B) two categories		
	a	b		a	b
VDG	0.85	0.87	VDG	0.80	0.83
R1	0.79	0.80	R1	0.76	0.77
R2	0.77	0.77	R2	0.70	0.69
R3	0.76	0.77	R3	0.68	0.70
R4	0.82	0.82	R4	0.77	0.78
RG	0.77	0.78	RG	0.70	0.72
average reader	0.79	0.79	average reader	0.73	0.74

Table 3.2: Mean inter-exam agreement kappa, based on four (A) and two (B) categories, using (a) all pairs (N=500), and (b) pairs with a screening interval of 20-28 months (N=373). Next to the individual readers we give the average kappa value and the kappa for VDG.

	VDG	R1	R2	R3	R4
VDG		0.70	0.63	0.65	0.71
R1	**0.78**		0.73	0.77	0.78
R2	**0.73**	**0.78**		0.74	0.74
R3	**0.76**	**0.79**	**0.78**		0.74
R4	**0.77**	**0.83**	**0.78**	**0.80**	

Table 3.3: Inter-reader agreement kappa, based on two (normal font) and four categories (bold font).

	current assessment VDG							
prior assessment VDG	1		2		3		4	
1	85	(17.0)	9	(1.8)	0	(0.0)	0	(0.0)
2	53	(10.6)	135	(27.0)	14	(2.8)	0	(0.0)
3	0	(0.0)	34	(6.8)	120	(24.0)	6	(1.2)
4	0	(0.0)	0	(0.0)	10	(2.0)	34	(6.8)

	current assessment R1							
prior assessment R1	1		2		3		4	
1	103	(20.6)	17	(3.4)	0	(0.0)	0	(0.0)
2	62	(12.4)	123	(24.6)	7	(1.4)	0	(0.0)
3	1	(0.2)	46	(9.2)	115	(23.0)	6	(1.2)
4	1	(0.2)	0	(0.0)	5	(1.0)	14	(2.8)

	current assessment R2							
prior assessment R2	1		2		3		4	
1	60	(12.0)	13	(2.6)	0	(0.0)	0	(0.0)
2	33	(6.6)	231	(46.2)	26	(5.2)	0	(0.0)
3	1	(0.2)	32	(6.4)	76	(15.2)	11	(2.2)
4	0	(0.0)	0	(0.0)	4	(0.8)	13	(2.6)

	current assessment R3							
prior assessment R3	1		2		3		4	
1	68	(13.6)	21	(4.2)	0	(0.0)	1	(0.2)
2	63	(12.6)	154	(30.8)	16	(3.2)	0	(0.0)
3	0	(0.0)	51	(10.2)	64	(12.8)	10	(2.0)
4	0	(0.0)	1	(0.2)	17	(3.4)	34	(6.8)

	current assessment R4							
prior assessment R4	1		2		3		4	
1	97	(19.4)	27	(5.4)	0	(0.0)	0	(0.0)
2	35	(7.0)	134	(26.8)	14	(2.8)	0	(0.0)
3	1	(0.2)	39	(7.8)	93	(18.6)	17	(3.4)
4	0	(0.0)	0	(0.0)	18	(3.6)	25	(5.0)

Table 3.4: Contingency tables of breast density at prior and current screening of the Volpara density grade (VDG) and the four readers. Given is the number of pairs and the percentage.

3.3.3 Intra-reader agreement and the effect of processing

The intra-reader agreement was calculated based on 50 exams for each processing (IRA_a and IRA_b). For two readers the breast density score was more consistent when images were read twice with the original processing, while two readers had a higher agreement for images presented with in-house processing. The kappa values are given in Table 3.5. Also kappa values for the 50 exams that were read with mixed processing, in session 1 or 2 with original processing and in session 3 or 4 with in-house processing, are shown in Table 3.5. When comparing the four agreements of each reader with each other, no significant differences were found.

	A) four categories					B) two categories			
	IRA_a	IRA_b	IRA_c	IRA_d		IRA_a	IRA_b	IRA_c	IRA_d
R1	0.72	0.73	0.82	0.74	R1	0.70	0.64	0.69	0.65
R2	0.87	0.84	0.72	0.77	R2	0.77	0.81	0.58	0.64
R3	0.82	0.77	0.84	0.80	R3	0.63	0.71	0.76	0.62
R4	0.86	0.90	0.84	0.83	R4	0.84	0.86	0.77	0.82

Table 3.5: Intra-reader agreement kappa, based on 50 exams in each category, see Figure 3.2. IRA_a is based on in-house processing and IRA_b is based on only original processing. For the combination of both processings (once read with original processing and once read with in-house processing) IRA_c and IRA_d are displayed. The agreements for four categories (A) and two categories (B) are determined.

3.4 Discussion

In this study, we compared the consistency of breast density assessment in serial screening mammograms obtained with the fully automated volumetric density measurements of Volpara to the consistency of human readers who used the BI-RADS categories. We made the comparison between human readers and automated classification using two categorisations. We first used a four point scale, as there are four BI-RADS and VDG classes. Additionally, we grouped the classes into a non-dense and a dense category. The latter may be more relevant in practice when density measurements are used to determine whether or not the mammography exam is sufficient for an individual woman and if supplemental screening should be offered.

We found that more women stay in the same category over time when the classification was done with VDG than with the simulated group reading (p-value 0.038) which best represents screening practice. Using two classes, VDG gave the same score to both exams in 90.4% of the cases. An agreement between 86.2% and 89.2% was found for the human readers. Furthermore, we found that the kappa value was significantly higher with VDG compared to the kappa value of the group reading ($p<0.001$ and $p=0.010$ for two and four

classes). As the automated density measurement gives more consistent results, especially when compared to the simulated group reading, this method may be preferred over human readers for breast density assessment.

We assume that breast density categories assigned to current and prior exams of the same woman should generally be similar. However, some differences are expected, because there is a gradual decrease in density with an increase of lifetime [25]. Therefore, we expect to observe some women changing from a higher to a lower density category, while an increase in density should occur less often. Our data confirmed this expectation. Deviations from the normal pattern may occur due to weight loss or the use of hormone replacement therapy (HRT), which may increase density. It should be noted, however, that HRT is not frequently used in the Netherlands [86]. In 2004, 3.39% of the women aged 40-74 made use of HRT. As the causes for an increase of density in postmenopausal women are limited, we see measurement variations as the main cause for a category change into a denser category.

The breast density assessment of the current mammogram was independent of the assessment of the prior exam. This may not reflect future screening practice. When density is used for stratification in screening, radiologists will know the density score given to the prior exam. This knowledge may influence the assessment of the current exam, leading to a higher agreement between the two density assessments. In this way, only significant changes will lead to a change in density category. Such a strategy can be implemented using automated scores as well.

To compute volumetric density, both the absolute fibroglandular tissue volume and breast volume are calculated. Additional comparisons of fibroglandular tissue volume and breast volume could be used to judge whether a category change was caused by a strong change of fibroglandular tissue volume or by a change of breast volume caused by weight loss or gain. With VDG both parameters can contribute to a category change from non-dense to dense [87]. Other than with VDG, stratification into different screening regimes could be done with fibroglandular tissue volume directly or with a combination of fibroglandular tissue volume and percent density.

Next to the inter-exam agreement, we measured the inter-reader agreement. The kappa values found were comparable to the values found in literature [40, 41]. This confirms that an inter-reader variability regarding the BI-RADS categories exists. Furthermore, we found a higher agreement between readers than between readers and Volpara. This finding might suggest that the readers capture masking and breast density in a different way than the software and that the software might miss some of the masking and risk associations. The study of Eng [24] however showed, that both BI-RADS and Volpara correlated with breast cancer risk, and that the association was stronger for Volpara than for BI-RADS.

To obtain the VDG categories, percent density was averaged over all possible images of an exam. Therefore, the variance on the percent density measurement might be smaller with MLO and CC views, compared to exams with only MLO views. However, volumetric

breast density estimates in the CC views are usually somewhat higher than the estimates based on MLO views [88], because the breast volume imaged in CC views is generally smaller. Taking the CC views into account could therefore lead to a slightly higher average percent density estimate and occasionally to a higher density category. As both human readers and the computer work with the same data, we do not believe that the limited availability of CC views has affected our results.

A potential limiting factor in our study is that the mammograms were presented to the readers with different types of processing. Therefore, we investigated if there were systematic effects on density classification related to processing. We compared the intra-reader agreement on images that were read with different types of processing and found that there was no significant difference. Therefore, we conclude that the type of processing plays only a minor role. It is remarked that in screening practice exams may be acquired with systems from different vendors, which means that different processing methods will occur also in practice.

Variations in breast density categories in serial screening mammograms have previously been studied by Spayne [89], Harvey [90] and Singh [91], see also Table 3.6. Spayne made use of film mammograms (screening interval 3-24 months) and found agreement in BI-RADS categories in 77.2% and 87.4% of their cases using four and two classes, respectively. In the study by Harvey an agreement of 69.8% and 83.54% was found for digital mammography, respectively. Categories were assigned in clinical practice and in only 19.8% of the pairs both exams were scored by the same radiologist. A higher percent agreement for the majority of readers was found in our study, using two categories, where digital mammograms were used. Singh studied the inter-exam agreement with three readers and another automated software system for volumetric breast density measurement (Quantra, Hologic, Bedford, USA) by calculating intraclass correlation coefficients. The results are comparable to the results found in this study, even though different study populations were used. The median interval between the two exams of the 144 pairs used by Singh was 13.2 months and is therefore much smaller than the interval used in this study. As there was less time between the exams than in our study, it is likely that in Singh's study density changes were smaller. Furthermore, in that study all prior and current mammograms were read sequentially with an interval of four weeks, while we randomised the order of priors and currents and had a minimum interval of only one week. However, we do not think that the short period between the readings used in our study led to an increased agreement due to a memory effect. Readers rapidly reviewed the cases and it is unlikely that they did remember their scores. On the other hand, it may well be that variability of their criteria for the categorisation increases with the interval length, which would cause a decrease of agreement over time. In that regard, in screening practice the reader agreement might be lower than we found, because the screening interval is in reality much longer than the interval in our experiment.

To conclude, with automated volumetric breast density measurements a more consistent density assessment of serial screening mammograms was observed than with the density assessment performed by trained clinicians. The use of volumetric breast density software led to a significantly higher consistency than the group reading, where mammograms were distributed randomly over the readers to mimic screening practice. Given the high consistency and the lack of potential inter- and intra-reader variability, automated density assessment may be preferred for breast density stratified screening.

	Spayne et al. [89]	Harvey et al. [90]	Singh et al. [91]	Our study
Type of acquisition	Film	FFDM	FFDM	FFDM
Study population	Breast Cancer Screening Program	Breast Cancer Screening Program	Hospital	Breast Cancer Screening Program
Density assessment	BI-RADS (combination of 3rd and 4th edition)	BI-RADS (combination of 3rd and 4th edition)	BI-RADS (4th edition), Quantra	BI-RADS (4th edition), Volpara
Number of mammogram pairs	11,755	87,066	141	500
Mean interval between mammograms (months)	13.1	14.3	13.2	30.0
BI-RADS assessment	Clinical practice	Clinical practice	Reader study (3 readers)	Reader study (4 readers)
Percent agreement BI-RADS categories	77.2%	69.8%	unknown	64-76%
Kappa agreement of BI-RADS categories	0.58 simple kappa, 0.70 weighted kappa	0.54	unknown	0.76-0.82 weighted kappa

Table 3.6: Overview of studies in which breast density of serial mammograms was investigated.

4

Volumetric breast density estimation in digital mammograms

Original title: Optimization of volumetric breast density estimation in digital mammograms.

K. Holland, A. Gubern-Mérida, R.M. Mann and N. Karssemeijer

Published in: *Physics in Medicine and Biology*, 2017, 62:3779-3797

Abstract

Fibroglandular tissue volume and percent density can be estimated in unprocessed mammograms using a physics-based method, which relies on an internal reference value representing the projection of fat only. However, pixels representing fat only may not be present in dense breasts, causing an underestimation of density measurements. In this work, we investigate alternative approaches for obtaining a tissue reference value to improve density estimations, particularly in dense breasts.

Two of three investigated reference values ($F1, F2$) are percentiles of the pixel value distribution in the breast interior (the contact area of breast and compression paddle). $F1$ is determined in a small breast interior, which minimises the risk that peripheral pixels are included in the measurement at the cost of increasing the chance that no proper reference can be found. $F2$ is obtained using a larger breast interior. The new approach which is developed for very dense breasts does not require the presence of a fatty tissue region. As reference region we select the densest region in the mammogram and assume that this represents a projection of entirely dense tissue embedded between the subcutaneous fatty tissue layers. By measuring the thickness of the fat layers a reference ($F3$) can be computed. To obtain accurate breast density estimates irrespective of breast composition we investigated a combination of the results of the three reference values. We collected 202 pairs of MRI's and digital mammograms from 119 women. We compared the percent dense volume estimates based on both modalities and calculated Pearson's correlation coefficients.

With the references $F1 - F3$ we found, respectively, a correlation of R=0.80, R=0.89 and R=0.74. Best results were obtained with the combination of the density estimations (R=0.90).

Results show that better volumetric density estimates can be obtained with the hybrid method, in particular for dense breasts, when algorithms are combined to obtain a fatty tissue reference value depending on breast composition.

4.1 Introduction

Mammographic breast density has been identified as an important risk factor for developing breast cancer [10, 22, 24]. Eng et al. [24] showed that the breast cancer risk is highly correlated to automated breast density measurements. Breast density was assessed in four density categories, and the breast cancer risk was up to 8.26 times higher for women with dense breasts (category 4) compared to women with non-dense breasts (category 1). Furthermore, it has been shown that sensitivity of mammography decreases with an increase of breast density [26, 27, 29, 31, 92]. Low sensitivity of mammography in women with dense breasts is explained by the fact that a tumour can be masked by the presence of fibroglandular tissue (also referred to as dense tissue), which is a combination of connective tissue structures and epithelial tissue. Fibroglandular and cancerous tissues have a similar attenuation for X-rays, and thus appear equally white in the mammogram while fatty tissue appears almost transparent.

Breast cancer screening programs are established in many countries. Starting at an age of 40-50 years, these programs offer women periodical breast cancer screening exams with mammography. Personalised breast cancer screening depending on familial risk and breast density, involving other modalities such as ultrasound or MRI, is under discussion [33, 34, 93]. Screening with MRI is already recommended for women with a lifetime risk of more than 20-25% [94, 95]. To implement personalised screening based on density, accurate and objective methods to estimate breast density need to be available.

Different methods have been developed to estimate the amount of fibroglandular tissue in the breast. An overview of automatic mammographic density segmentation techniques has been published recently [54]. The methods can be categorised in two types: area-based and volume-based methods. Area-based methods [55, 96–98] were developed to objectively reproduce the density categories described in the Breast Imaging Reporting and Data System (BI-RADS) [37]. The BI-RADS density grade is a four-point scale used by radiologists to estimate the percentage area of fibroglandular tissue that is projected in the mammogram. A major limitation of area-based methods is that the 3D structure of the breast is not taken into account as only the projection of fibroglandular tissue is represented. Therefore, the resulting percentage dense area is not invariant to compression and projection angle. Some area based methods also have the disadvantage that human interaction is required [55].

To overcome these problems, several methods have been proposed to fully automatically estimate volumetric percent density, defined as fibroglandular tissue volume divided by breast volume, on digital mammograms (DM) [45–47, 78, 79, 99]. These methods employ a physics based model of the X-ray image acquisition process and assume that the breast is composed of two types of tissue: fat and fibroglandular tissue. Eng et al. have shown that these breast density estimations correlate well with each other and with breast cancer risk. One of the algorithms used by Eng et al. is Volpara (Volpara Health Technologies, Wellington, New Zealand) which is based on the work of Highnam [45, 47] and van Engeland [46].

The methods described in Highnam and van Engeland make use of an internal calibration with a pixel that ideally belongs to the projection of fat only. Only pixels in the breast interior are taken into account when estimating the internal reference value, where the interior is defined as the region where the breast is in contact with the compression paddle, excluding the peripheral region of the mammogram in which the breast thickness is smaller. An important limitation of this calibration approach is that the pixel value used as reference may not be representative for only fatty tissue in the breast interior if the breast is very dense. This leads to an underestimation of the fibroglandular tissue volume and percent density in dense breasts [51–53].
The aim of our study is to improve breast density estimation in very dense breasts and to find an approach that gives reliable breast density estimations in all types of breasts. We compared the use of alternative methods for obtaining a fatty tissue reference value. For that purpose, we estimated the fibroglandular tissue volume and percent density with three different reference values on mammography and we compared these estimates to reference breast density estimations obtained from MRI data. We first used a reference value obtained with a small breast interior to avoid that pixels of the periphery are accidentally included. The second method used an enlarged breast interior, which is more likely to include a pixel value that represents the projection of only fatty tissue, but has a larger risk to overestimate the reference value due to inclusion of pixels of the peripheral zone. The third approach is based on the idea that also pixel values representing the largest proportion of fibroglandular tissue may serve as a reference. By estimating the amount of fibroglandular tissue that corresponds to the densest part in the mammogram e.g. the difference between the breast thickness and the amount of subcutaneous fat, the fatty tissue reference value is computed. The aim of the latter method is to deal with very dense breasts. Obviously, this method will fail when the breast is not very dense. To overcome this drawback, we investigated if a combination of the three breast density measurements can lead to an overall improvement of breast density estimation results.

4.2 Methods

4.2.1 Preprocessing

All mammograms underwent some steps of preprocessing. First the images were segmented into breast tissue, pectoral muscle (in case of a mediolateral oblique - MLO view) and background [100]. Additionally, we performed a thickness correction of the peripheral region. We also determined the Euclidean distance $d(r)$ from each pixel location to the skinline in the mammographic projection. The distance $d(r)$ is used several times throughout the paper and is needed to define the breast interior.

4.2.2 Fibroglandular tissue volume

As described in van Engeland et al. [46], the volume of fibroglandular tissue can be computed from unprocessed (raw) digital mammograms based on a physical model of image acquisition and on the assumption that the breast is composed of two types of tissue: dense tissue and fatty tissue. The dense tissue thickness h_{d} at each pixel location can be calculated with the following formula:

$$h_{\mathrm{d}}(r) = -\frac{1}{\mu_{\mathrm{d,eff}} - \mu_{\mathrm{f,eff}}} \ln\left(\frac{g(r)}{F}\right) \tag{4.1}$$

where, $g(r)$ is the pixel value at position r and F is the pixel value in a fatty tissue reference region where h_{d} is supposed to be zero. The effective attenuation coefficients, $\mu_{\mathrm{f,eff}}$ and $\mu_{\mathrm{d,eff}}$ for fatty and dense tissue, respectively, vary with the breast thickness and with the anode / filter combination of the X-ray tube and are described in van Engeland et al.. The total dense tissue volume (VDT) is given by the integral over the projected breast area B:

$$VDT = \int_B h_{\mathrm{d}}(r) d^2r = -\frac{1}{\mu_{\mathrm{d,eff}} - \mu_{\mathrm{f,eff}}} \int_B \ln\left(\frac{g(r)}{F}\right) d^2r \tag{4.2}$$

4.2.3 Fatty tissue reference value

In this work, we studied three approaches to obtain an estimate for the pixel value that corresponds to the projection of fatty tissue only. Two reference values ($F1$ and $F2$) are based on the pixel value distribution in the breast interior and the third reference value ($F3$) was calculated by estimating the proportion of dense tissue in the densest location in the breast. Each approach is explained in the following section.

Reference value $F1$ - the maximum pixel value in a small breast interior region

To obtain the reference value $F1$, we used the same approach as described in van Engeland et al. The pixel value representative for fatty tissue is determined by taking a large quantile (0.99) of the pixel value histogram computed in the breast interior. We used this approach instead of taking the maximum pixel value, because large pixel values may appear in the mammogram due to artefacts or noise. With this approach the breast interior is rather small and depends on the maximum Euclidean distance computed from the breast pixels to the skin-line. To determine this distance, we first calculate the Euclidean distance from all breast pixels (the pixels that are enclosed by the pectoral muscle boundary, the skin-line and the image edge) to the skin-line. Of these distances we take the maximum. In most cases, the point with the maximum distance to the skin is located on the pectoral muscle or breast boundary, but it is noted that depending on the shape of the breast this point may also be located more inward from the boundary. The interior is then defined as the breast pixels that have at least a distance of 0.4 times the maximum Euclidean distance to the skin-line. An example is shown in Figure 4.1B.

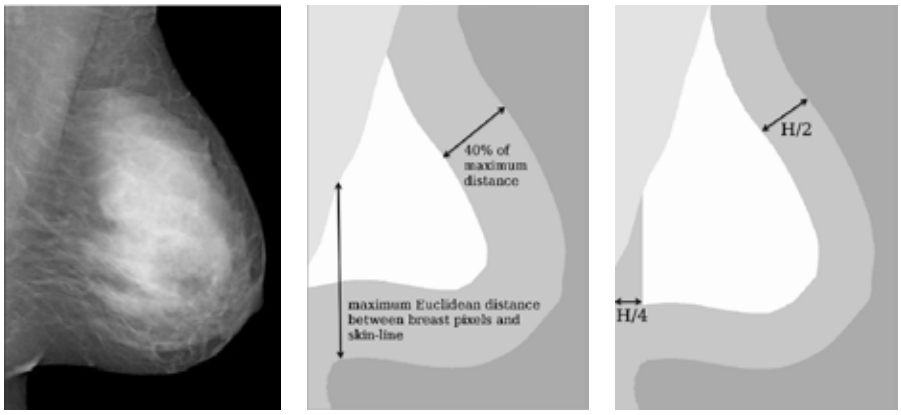


Figure 4.1: A processed mammogram (A) and the segmentation of breast tissue, pectoral muscle and background (B and C). A graphical representation of the small (B) and large (C) breast interior are also shown. The breast interior as defined in B) is used to compute $F1$. The breast interior shown in C) is used for the computation of $F2$ and $F3$.

Reference value $F2$ - the maximum pixel value in a large breast interior region

In denser breasts, the interior of the breast as defined previously might not contain a pixel representing only fatty tissue. Therefore, we used a second definition of the breast interior region in which pixels closer to the skin-line and located within the fully compressed area are also included.

We defined the larger breast interior as the pixels that have a minimum distance to the skin-line of half the compressed breast thickness (H). Additionally, we excluded pixels that are too close to the vertical image edge, as this region is sometimes not representative due to poor compression, in particular in MLO views. Furthermore, we want to prevent that the pectoral muscle is part of the breast interior in case it is visible in a craniocaudal (CC) view. The breast interior pixels have a minimum distance of a quarter of the breast thickness to the vertical image edge. The breast interior region is given for one mammogram in Figure 4.1C. To compute the reference value, we again used a large quantile (0.99) of the pixel value histogram. We call this reference value $F2$.

Reference value $F3$ - an estimate from the densest region (minimum pixel value of the mammogram)

When the breast is very dense, the methods above become inaccurate, as it becomes unlikely that the obtained reference value is representative for the projection of only fatty tissue. In this section, we propose a method to obtain a suitable reference value using the densest region in the mammogram. Given the pixel value corresponding to a dense pixel (g_{dense}) at a certain location r, it is possible to compute the fatty tissue reference value by estimation

of the corresponding thickness of dense tissue $h_{\text{dense}}(r)$. This can be seen if we rewrite formula 4.1:

$$\begin{aligned} F &= \frac{g(r)}{\exp\left(-(\mu_{\text{d,eff}} - \mu_{\text{f,eff}})h_{\text{d}}(r)\right)} \\ F3 &= \frac{g_{\text{dense}}}{\exp\left(-(\mu_{\text{d,eff}} - \mu_{\text{f,eff}})h_{\text{dense}}\right)} \end{aligned} \tag{4.3}$$

The reference value computed in this way is denoted by $F3$. Note that there is no reference region in the mammogram with this pixel value, but that it can be derived when we can estimate h_{dense} at the location of g_{dense}. To find the reference pixel value g_{dense}, we make use of the large breast interior as used for $F2$ (see Figure 4.1C) and determine the minimum using the 0.01 quantile of the pixel value distribution.

To obtain h_{dense} we have to estimate the amount of fibroglandular tissue that corresponds to the pixel value g_{dense}. Since the reference value g_{dense} corresponds to the densest region in the mammogram, we assume that it represents the projection of dense tissue and the layers of subcutaneous fat. Therefore, this pixel value is also representative for the maximum dense tissue fraction ($MDTF$) in the compressed breast projection, i.e. the maximum thickness of dense tissue in the path of the X-ray beam divided by the thickness of the compressed breast. Based on this idea h_{dense} is estimated as $MDTF \times H$ with H the compressed breast thickness. To estimate h_{dense}, the $MDTF$ in the direction of the X-ray beam is needed. As this information is unavailable, we calculate the $MDTF$ in the image plane and assume that this value is representative for the $MDTF$ in the direction of the X-ray beam. To make this approach work, two key assumptions need to be fulfilled: 1) in a first approximation the $MDTF$ is direction independent with respect to a rotation around the anterior-posterior axis and 2) that the $MDTF$ does not change when the breast is compressed. In section 4.3.5 we investigate to what extent these assumptions are valid by using MRI data. The $MDTF$ in the image plane is calculated by constructing paths that represent plausible X-ray trajectories through the breast when it would be decompressed, rotated by 90 degrees, and compressed again. We used slightly curved instead of linear paths, to take into account that when imaging a compressed breast the X-ray beam will be approximately perpendicular to the skin and the subcutaneous fat layer.

The curved paths are defined by taking points on a circle centred at a point (P) outside of the breast. A schematic overview is shown in Figure 4.2. To define this point P, we calculate the minimum distance between the nipple (N) and the pectoral muscle (M). In CC views, we used the vertical image edge instead of the pectoral boundary. The point P has a distance to the nipple of two times the distance NM and lies on the straight line with the points N and M. The nipple is assumed to be located on the skin-line at the position with the maximum distance to the pectoral muscle or the vertical image edge for MLO views and CC views, respectively.

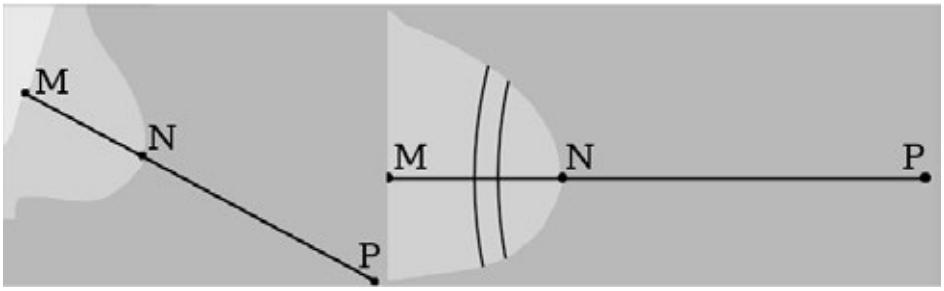


Figure 4.2: A schematic view of the land marks nipple (N), closest point (M) of the nipple to the pectoral muscle / image edge, for MLO and CC views, respectively, and the new point P. In the example of the CC view (right side) are also two curved paths given. All points of a counting path have the same distance to the point P.

To estimate $MDTF$ in the plane of the mammogram we first classify all pixels as dense or non-dense with a Random Forest classifier on a rescaled 200 μm image, using features similar to the ones used in Kallenberg et al. [101]. Then, for each path, we count the number of dense and non-dense pixels. To minimise variation we group paths to bands with a width of 0.2 mm with a sliding window approach and calculate the dense tissue fraction of these bands. The $MDTF$ of a mammogram is the absolute maximum of the dense tissue fractions. To estimate h_{dense}, we use the $MDTF$ averaged over all mammograms of an examination (left and right breast, and MLO and CC view), to minimise variation. The $MDTF$ of the MLO and the CC view are assumed to be similar due to the rotation invariance of the MDTF, while the $MDTF$ of the left and right breast are assumed to be similar as the amount of breast density and the breast density distribution of the left and right breast are comparable to each other. Only paths with a minimum distance of $NP + H/8$ are included, to exclude paths that are too close to the nipple. Furthermore, we exclude paths that are too close to the pectoral muscle to prevent that inaccuracies in the pectoral muscle segmentation influence the result. We also use a distance of $H/8$ here. Using the thickness of the compressed breast H we estimate h_{dense} by computing $MDTF \times H$.

4.2.4 Combination of fibroglandular tissue volume estimations

Previous studies showed that the breast density estimates obtained with $F1$ provide accurate results for non-dense breasts. However, the reference values $F2$ and $F3$ were obtained under the assumption that the breast is dense. Hence, it is likely that the breast density estimations obtained with these reference values do not give accurate results in non-dense breasts. In this study, we developed two hybrid approaches that combine the breast density estimations to obtain suitable results for the complete range of breast densities.

The first combination scheme takes only into account results from $F1$ and $F2$. Based on the $MDTF$ a threshold t is set. If $MDTF$ is below that threshold, the fibroglandular tissue volume estimate obtained with $F1$ is used. Otherwise we use results from reference $F2$.

The second combination scheme of the breast density estimates involves three parameters (t, a and b) and results from the three reference values are combined. The first parameter is a threshold (t). If $MDTF$ is below that threshold, the fibroglandular tissue volume estimate obtained with $F1$ is used. In case of a $MDTF$ above that threshold, a linear combination of the other two fibroglandular tissue volume estimations is used, where the weights depend on the $MDTF$. The combination scheme can be written as

$$VDT = \begin{cases} VDT1 & \text{if } MDTF < t \\ (1-w) \times VDT2 + w \times VDT3 & \text{if } MDTF \geq t \end{cases} \tag{4.4}$$

with $w = a \times MDTF + b$ and $VDT1$, $VDT2$ and $VDT3$ the dense tissue volumes obtained with the reference values $F1$, $F2$ and $F3$, respectively.
In the evaluation, five fold cross validation was used to obtain optimal values for the parameters. The data set was divided into five groups, or folds, of equal size. To avoid bias all exams of a woman were in the same fold. Then, four folds were used to optimise the parameters, which were then applied on the fifth fold. This process was repeated such that the obtained parameters were once applied to each fold. The parameters were obtained by minimising the sum of the squared residuals in the comparison of percent density between mammography and MRI averaged per breast.

4.2.5 Breast volume

In this work, we used two approaches to estimate the breast volume. The first one is based on the semi-circle model [102]. In that approach, which was applied in van Engeland, the interior is assumed to consist of parallel planes, while the peripheral zone is approximated by semi-circles. The thickness as a function of the compressed breast thickness (H) and the Euclidean distance between the breast pixel location and the skin-line ($d(r)$) is given with the following formula:

$$h(r) = \begin{cases} 2\left[(H/2)^2 - ((H/2) - d(r))^2\right]^{1/2} & \text{if } d(r) < H/2 \\ H & \text{if } d(r) \geq H/2 \end{cases} \tag{4.5}$$

The second breast volume estimate algorithm used in this study is based on the work of de Groot et al. [103]. That method uses a model where the peripheral zone of the breast in the lateral direction, i.e. the region where tissue is not in contact with the compression paddle, is assumed to be have half the width of the peripheral zone in the posterior-anterior direction, due to forces related to the attachment of the breast to the chest wall.
To segment the peripheral zone, de Groot performs a sequence of scaling with a factor of two along the rows, erosion, dilation and rescaling of the segmented mammogram. The pixels in the lateral direction that are closer to the skin-line than H/4 are considered to belong to the peripheral zone, while in the posterior-anterior direction pixels closer than

H/2 are included. The different steps are shown in Figure 4.3.
Since the description proposed by de Groot et al. is for CC mammograms only, we adapted this approach for MLO views. We introduced a coordinate system in which the y-axis is defined by the pectoral muscle boundary, and the x-axis is perpendicular to the y-axis and goes through the nipple position. For each point within the breast segmentation, we determined the absolute angle ($|\alpha|$) between the x-axis and the line going through that point and the origin. Points that were closer to the skin-line than

$$\frac{H}{2}\left(1-\frac{|\alpha|}{\pi}\right) \tag{4.6}$$

were considered to belong to the region that is not in contact with the compression plate. Figure 4.4 shows a MLO mammogram with the coordinate system. Like in de Groot, we used H, obtained from the DICOM header, as thickness for the fully compressed breast region, and a thickness of $H\pi/4$ (average thickness of the peripheral zone) for the peripheral zone. The breast volume is the integral over the breast thickness over the breast area.

4.2.6 Segmentation of MR images

For evaluation of the methods we used breast MRI scans. First, the MRI scans were segmented into breast tissue and background. Each breast voxel was then labelled as fibroglandular tissue or fatty tissue. For most MRI's this was performed automatically with a previously developed method [104]. Some images were segmented manually. A manual segmentation was necessary, when the automatic segmentation was not accurate enough judged by a radiologist. Manual segmentations were performed by a trained researcher and were reviewed by a radiologist with expertise in breast MRI.
Manual segmentation was done as follows: In the axial view, in every 5-10 slices a contour was drawn around the breast outline using spline interpolation. Using these contours, the breast outline in the remaining slices was computed using a spline surface function to get the breast mask. These masks were checked and in case of an inaccurate segmentation, additional contours were drawn until the masks were accurate enough. Within the breast mask, contours for the segmentation into fibroglandular and fatty tissue were drawn in the same slices as the contours to segment the breast from background. These contours were extended to the whole breast with the method mentioned above. Within this fibroglandular tissue mask a threshold was applied to distinguish fibroglandular tissue from fatty tissue. Voxels outside the fibroglandular tissue mask but inside the breast mask were considered to be fatty. Segmentations were performed on bias field corrected images. The N4 bias field correction algorithm [105] was used.

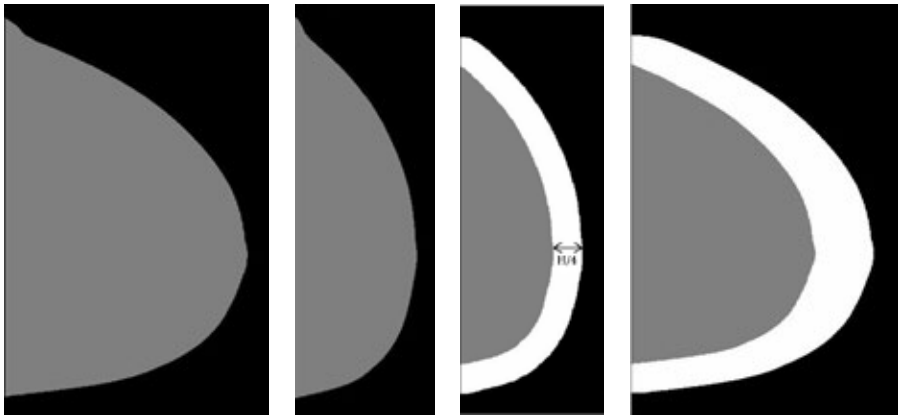


Figure 4.3: Image processing steps to determine the breast region that is in contact with the compression plate of a CC image. First, the segmented breast (A) is scaled to half width (B), then an erosion and dilation with a circle structure element with a diameter of $H/4$ is applied, leading to the band at the skin-line (C). That image is then rescaled to the original width (D).

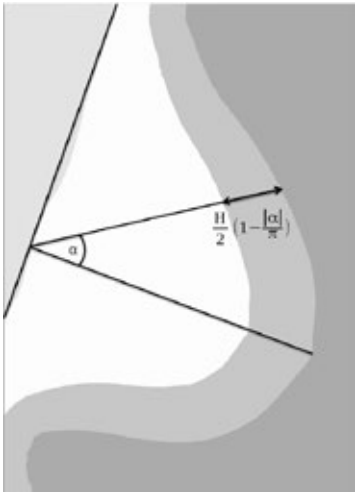


Figure 4.4: A schematic view of an MLO image and the differentiation into the fully compressed breast region (white) and the region where the breast is not in contact with the compression plate (grey). The coordinate system is based on the pectoral muscle segmentation and the estimated nipple position. For each point within the breast, the angle with the x-axis is determined (α). The two regions are defined by formula 4.6.

4.3 Experiments

4.3.1 Materials

We made use of 202 pairs of mammography and MR examinations of 119 women recorded between December 2000 and December 2011 in the Radboudumc. MRI exams and mammograms were performed within two months of each other. All digital mammograms were acquired on GE Senographe systems using standard clinical settings. For all examinations, a complete exam consisting of MLO and CC images of the left and right breast was available. We used the (raw) unprocessed data.

The MR examinations were performed on either a 1.5 or a 3 Tesla Siemens scanner (Magnetom Vision, Magnetom Avanto or Magnetom Trio), with a dedicated breast coil (CP Breast Array, Siemens, Erlangen, Germany). The segmentation of breast and fibroglandular tissue was performed on pre-contrast T1-weighted MR volumes without fat saturation. For 159 MRI examinations the automated segmentation was approved by a radiologist, 43 MRI's were segmented manually.

4.3.2 Comparison of mammography based fibroglandular tissue estimations to MRI data

The fibroglandular tissue volume estimations obtained with the three different reference values and the two combined approaches were compared to estimations based on MRI data. Results were averaged over MLO and CC views to obtain an estimate per breast. For an estimation per exam results were averaged over the left and right breast. Pearson's correlation coefficients and 95% confidence intervals (CI) were calculated using the statistical software package R. Because of the log-normal distribution of the data, correlation coefficients were computed after converting the measurements using the natural logarithm [52, 53].

4.3.3 Comparison of breast volumes based on mammography and MRI

Mammographic breast volumes obtained with the two different geometrical approaches, the semi-circle model and the adaption of de Groot et al., were compared to breast volumes based on MRI data. Results were averaged over both mammographic views and over the left and right breast to obtain a single score for each exam. The MRI based breast volumes of the left and right breast were averaged as well. Pearson's correlation coefficients and 95% CI were calculated. Based on the performance of the two algorithms, we decided to proceed with one algorithm for further computations.

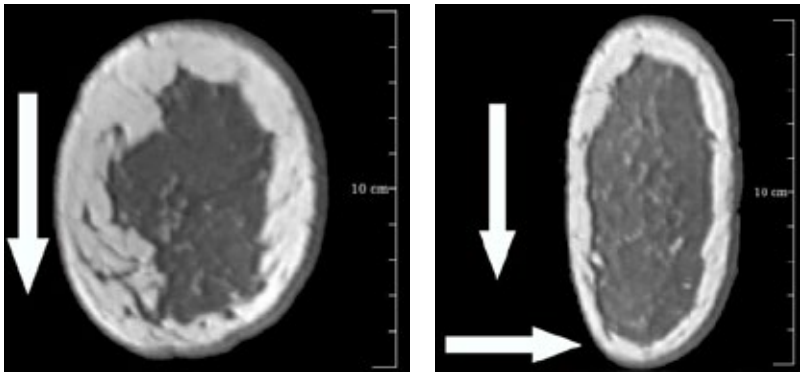
4.3.4 Evaluation of volumetric percent density estimations

The percent density estimations obtained with the three different reference values and the two combined estimates were compared to the estimations based on MRI data. Percent density is defined as fibroglandular tissue volume divided by breast volume. We made use of the breast volume estimations that are based on the work of de Groot et al. In the original work of van Engeland the semi-circle model was used to obtain the breast volume. For

comparison, percent density is given based on the fibroglandular tissue estimations with $F1$ and the breast volumes estimations of the semi-circle model as well. The estimations were again averaged over MLO and CC views to obtain an estimate per breast. To get an estimation for each exam, estimations were averaged over the left and right breast. Pearson's correlation coefficients and 95% CI were calculated.

4.3.5 Testing rotation invariance and effect of compression

To obtain a fatty tissue reference value with the dense tissue reference region approach we assume that the maximum dense tissue fraction ($MDTF$) in the direction of the mammographic compression (the direction of the X-ray beam) is the same as the $MDTF$ in the image plane. This assumption is based on the idea that the $MDTF$ does not depend on the direction of the measurement and is not affected by deformation of the breast due to compression. To investigate the validity of this idea we used the previously described data set of 202 MRI exams of 119 women. Additional, we included exams of 20 women who underwent MR guided biopsies and had breast density categorised as either BI-RADS 3 or 4. For each of these 20 women two MRI exams were available: An MRI recorded with the breast under compression in the MR guided biopsy procedure and a regular diagnostic MRI. The compression of the biopsied breast was always in the mediolateral direction. All MR images were manually segmented into fibroglandular tissue, fatty tissue and background.
To verify the rotation invariance of the $MDTF$ measurements, we took samples in regular (uncompressed) MRI exams along lines in multiple directions in the coronal planes. Samples in each direction were taken on a regular grid in a square region of $6 \times 6 cm^2$, were the centre of the square region projected to the centre of the breast. By doing so, we prevented taking samples close to the pectoral boundary or near the breast edge. Samples had a cross section of $2 \times 2 mm^2$. In each of these samples the fraction of dense tissue was computed from the segmented MRI. The $MDTF$ was computed as a function of the projection angle by taking the maximum dense tissue fraction in the square region while rotating the breast. We took samples using 36 different angles, ranging between 0 and 175 degrees in steps of five degrees, simulating 36 different projection directions. The mean $MDTF$ and the coefficient of variation when changing the projection angle were computed and displayed with a scatter plot. For this experiment we used both the set of 20 MRI's of uncompressed dense breasts and the 202 segmented MR images. In Figure 4.5A a coronal slice of a MRI volume is given. The sample direction is craniocaudal.
In a second experiment we looked at the effect of compression using the exams of the patients who underwent MR-guided breast biopsies. We determined the maximum dense tissue fraction in the MRI's of the compressed breasts in two directions, one in the compression direction and one perpendicular to it, both along lines in parallel with the chest wall. These correspond to the directions in which we compute the maximum dense tissue fractions in the mammograms, with the difference that the MRI's are compressed in



A B

Figure 4.5: Two MRI exams of the same patient: A) A coronal slice of an regular MRI volume, with an arrow indicating the craniocaudal direction in which the dense tissue fraction is determined, and B) a coronal slice of an MRI volume recorded in a MR guided biopsy procedure, in which the dense tissue fraction was determined in two directions: mediolateral and craniocaudal as indicated by vertical and horizontal arrows, respectively.

mediolateral directions while mammograms are compressed in the mediolateral oblique or craniocaudal direction. A visual example is shown in Figure 4.5B.

4.4 Results

4.4.1 Fibroglandular tissue volume

Figure 4.7 shows the results of the five methods for dense tissue volume estimation using MRI as a reference. The Pearson's correlation coefficients for the comparisons with MRI were 0.79, 0.80, 0.73, 0.83 and 0.86 when using $F1$, $F2$, $F3$, the combination of $F1$ and $F2$, and the combined estimates of $F1-F3$, respectively. In Table 4.1 the correlation coefficients are given for the estimations averaged per breast and per exam.

The parameters t, a and b that are needed to combine the results were obtained through cross validation. The optimal values for the parameters were determined in each fold and varied only a little. For both approaches, when combining $F1$ and $F2$, and when combining all three estimates, the optimal threshold was set to 0.35 (in four of the five folds). Therefore, the fibroglandular tissue volume as estimated with $F1$ is used if the $MDTF < 0.35$. The parameters a and b are needed when combining all three fibroglandular tissue volume estimates. In the majority of folds, the optimal values for a and b were 1.4 and -0.8, respectively. The $MDTF$ is a first indication for the density. So in non-dense breasts the estimates as obtained with $F1$ are used while for dense breasts a combination of $F2$ and $F3$ works best.

	per breast	per exam
$F1$	0.77 (0.73-0.81)	0.79 (0.73-0.84)
$F2$	0.78 (0.74-0.81)	0.80 (0.74-0.84)
$F3$	0.71 (0.66-0.75)	0.73 (0.66-0.79)
combination $F1$ and $F2$	0.81 (0.78-0.84)	0.83 (0.78-0.87)
combination $F1$, $F2$ and $F3$	0.83 (0.80-0.86)	0.86 (0.81-0.89)

Table 4.1: Pearson's correlation coefficients with 95% CI for the five fibroglandular tissue volume estimations.

4.4.2 Breast volume

Figure 4.6 shows results of the two breast volume estimation methods in comparison to MRI. Breast volume based on mammography has a linear relationship with the breast volume calculated from the MR images with both approaches. A Pearson's correlation coefficient of 0.96 (0.95-0.97 95%CI) was obtained for the semi-circle model and for the model in which the breast region that is in contact with the compression plates has a direction dependency. We decided to use the second approach, which is based on the idea of de Groot et al., for further calculations. This method appears to have less bias as the data points are closer to the identity line.

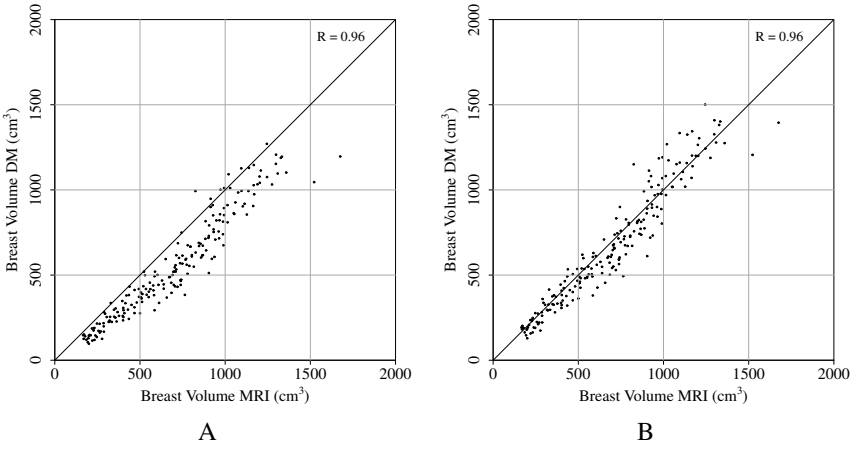
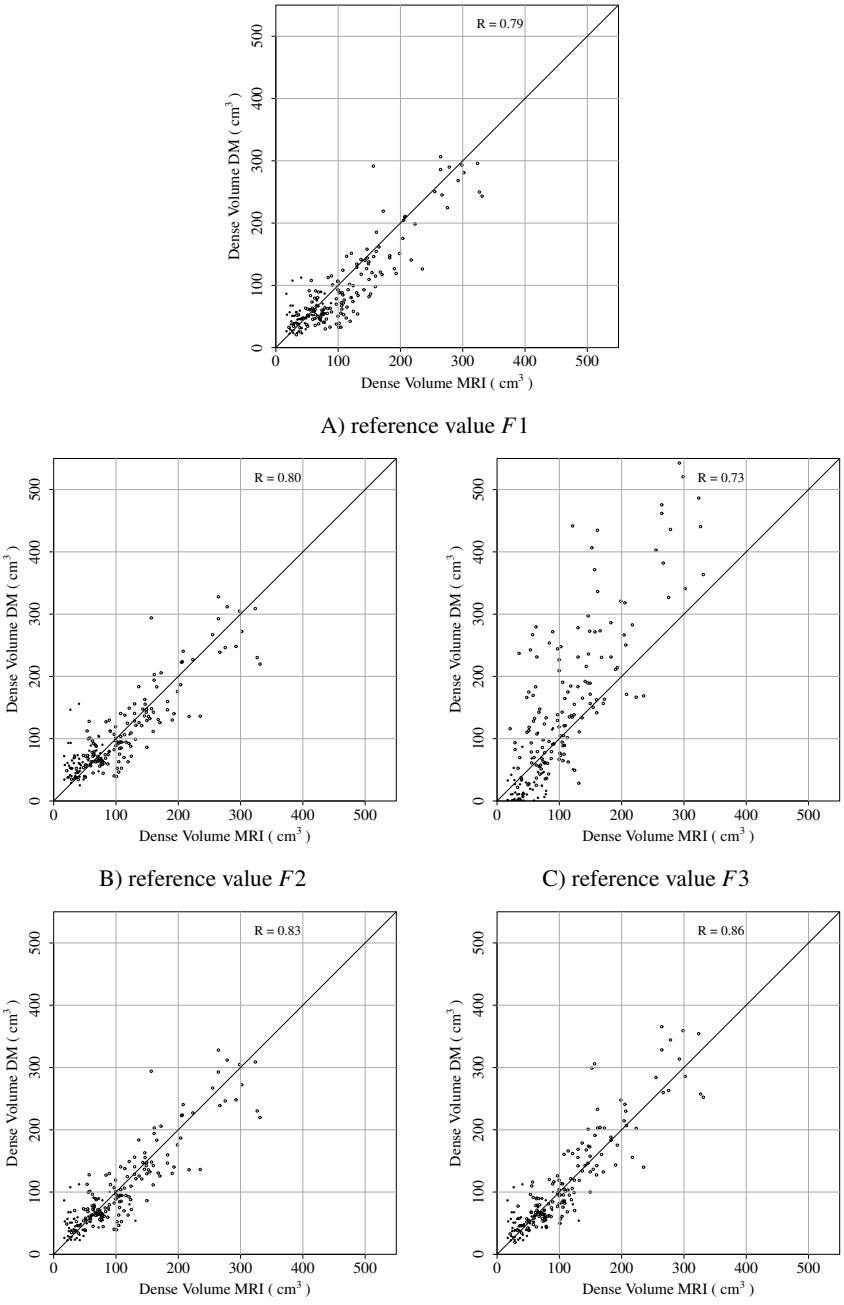


Figure 4.6: Comparisons of breast volume obtained with MRI and mammography, averaged per examination. In figure A) are the estimations with the semi-circle model, while in figure B) the estimates based on the idea of de Groot et al. are displayed, where the definition of the fully compressed breast region has a direction dependency.

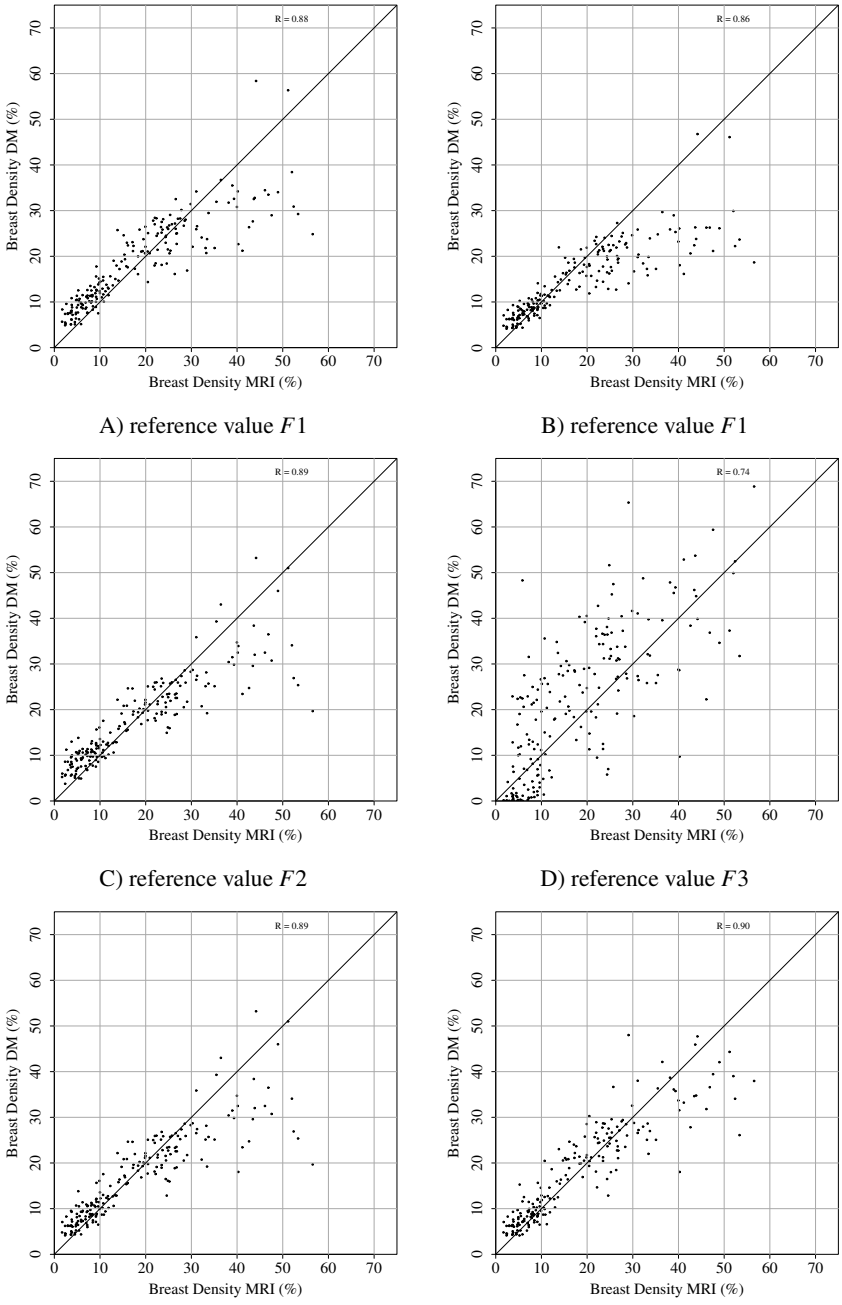


A) reference value $F1$

B) reference value $F2$ C) reference value $F3$

D) combination of reference value $F1$ & $F2$ E) combination of the three reference values

Figure 4.7: Comparison of fibroglandular tissue volume estimates per exam. In the subfigures A-C, the closed circles belong to cases with a $MDTF$ below 0.35, open circles have a $MDTF$ above 0.35. In the figures D-E, closed circles are used for exams that use the estimation of $F1$, while open circles are used when $F2$ (in D) or the combination of $F2$ and $F3$ (in E) is used.



A) reference value $F1$ B) reference value $F1$

C) reference value $F2$ D) reference value $F3$

E) combination of reference value $F1$ & $F2$ F) combination of the three reference values

Figure 4.8: Comparison of percent density estimations per exam. In A) the breast volume is estimated with the semi-circle model. For the results in B-F the breast volume as obtained from the second algorithm was used.

4.4.3 Volumetric breast density estimation

Figure 4.8 shows the comparison between volumetric percent density estimates from mammography and MRI. The Pearson's correlation coefficients between the volumetric percent density estimations per study were 0.86, 0.89, 0.74, 0.89 and 0.90 when using $F1$, $F2$, $F3$, the combination of $F1$ and $F2$ and the combination of $F1-F3$, respectively. When using the fibroglandular tissue estimations with $F1$ and the semi-circle model for the breast volume a correlation of 0.88 was found. Pearson's correlation coefficients with 95% CI are shown in Table 4.2 for comparisons per breast and per exam.

	per breast	per exam
$F1$ with semi-circle model	0.87 (0.84-0.89)	0.88 (0.84-0.91)
$F1$	0.84 (0.81-0.87)	0.86 (0.82-0.89)
$F2$	0.87 (0.85-0.89)	0.89 (0.86-0.91)
$F3$	0.73 (0.68-0.77)	0.74 (0.67-0.80)
combination $F1$ and $F2$	0.87 (0.84-0.89)	0.89 (0.85-0.91)
combination $F1$, $F2$ and $F3$	0.89 (0.87-0.91)	0.90 (0.88-0.93)

Table 4.2: Pearson's correlation coefficients with 95% CI for the six volumetric density estimations. When not indicated the breast volume as obtained with de Groot et al. was used.

4.4.4 Rotation invariance and the effect of breast deformation

We determined $MDTF$ values for different projection angles in the MR images of 20 dense breasts and the 202 segmented MR volumes. With a rotation interval of 5 degrees, this yielded to $MDTFs$ of 36 different angles. The mean $MDTF$ ranged between 0.11 and 0.95 and the coefficient of variation of the measurements ranged between 0.01 and 0.54. In Figure 4.9A a scatter plot of the mean $MDTF$ and the coefficient of variation is shown. We see that the coefficient of variation decreases with an increase of the mean $MDTF$. The low coefficient of variation of dense breasts (high mean $MDTF$), indicates that the $MDTF$ is almost direction invariant in extremely dense breasts. On the other hand, the coefficient of variation is larger of small mean $MDTF's$. As the $MDTF$ is a first indication for the breast density, we can see that the idea of direction invariance is violated in non-dense breasts. Hence, it is no surprise that the $MDTF$ used to estimate $F3$ becomes unreliable, resulting in unrealistic low dense volume measurements (see also Figure 4.7C).

The comparison of the maximum dense tissue fractions measured in coronal planes of the compressed breast MRI's is shown in Figure 4.9B. Results obtained in the mediolateral direction, the direction of the compression, are on the vertical axis, while results of the craniocaudal direction are plotted horizontally. It can be observed that the maximum dense tissue fraction ($MDTF$) in the two directions are quite similar, which supports the validity of our approach.

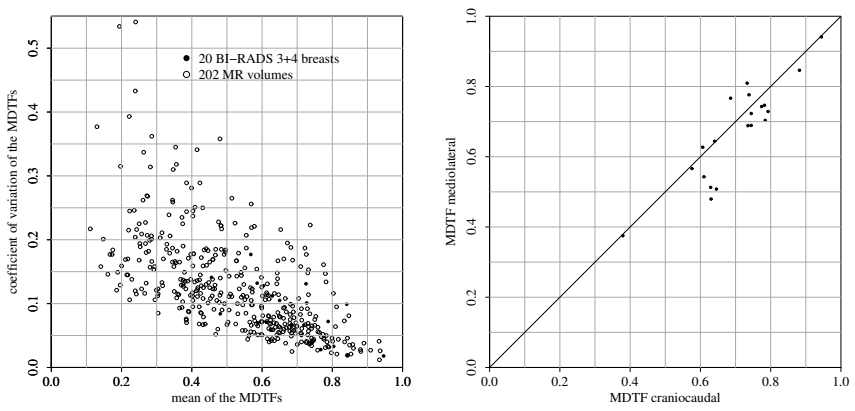


Figure 4.9: A) Mean and coefficient of variation of the 36 $MDTF$ measurements of the MR images of the 20 uncompressed breasts (closed circle) scored as BI-RADS density 3 or 4, and the 202 segmented MR images (open circle). B) Maximum dense tissue fraction measured in MRI's with compression of the breast in mediolateral direction. The dense tissue fractions in the direction of the compression are compared to the perpendicular direction.

4.5 Discussion

Accurate and reliable breast density measurements are needed when using density for stratification or breast cancer risk analysis. In this study, we experimented with the use of different internal reference values for calibration of a widely used volumetric breast density quantification method. We used reference values of pure fatty pixels using two different breast interior definitions and also proposed a novel approach, which employs the densest region in a mammogram as a reference. Hybrid approaches to obtain robust estimates in fatty and dense breasts, which combines breast density estimates, are also proposed.

We estimated the reference value of purely fatty tissue based on the pixel value distribution in the breast interior. The two reference values $F1$ and $F2$ use different definitions of the breast interior. The third reference value $F3$ was determined based on a pixel value representative for the densest region in the mammogram. When comparing volumetric percent density based on mammography to MRI data for each exam, Pearson's correlation coefficients of 0.86, 0.89 and 0.74 were found when using $F1$, $F2$ and $F3$, respectively, when using de Groot to obtain the breast volume estimate.

Results show that the individual performance of $F1$ and $F2$ is much better than the one of $F3$. Therefore, it might be questionable whether $F3$ adds a lot to the performance of a hybrid approach. However, in our experiments best results were obtained when combining all three estimations (Pearson's correlation 0.90). This can be explained by the fact that different breast density patterns require different approaches. Compared to the original method, a clear improvement is visible in the scatter plots when comparing the results to that of

the combined approach (Figure 4.8), in particular for dense breasts. The combination of results of $F1$ and $F2$ yielded a correlation coefficient of 0.89. Though the difference in correlation coefficient between the two hybrid methods is small, substantial differences are visible between the two scatter plots for dense breasts. This confirms that a new approach is necessary to deal with extremely dense breasts. The difference in correlation found in our study is not large, we expect larger differences when more cases with dense breasts would be included in the study set. An analysis with only dense breasts might be carried out in the future.

The novel method based on using the densest region in a mammogram as a reference relies heavily on the assumption that the maximum dense tissue fraction ($MDTF$) does not strongly depend on the projection angle and the compression. We tested the assumption that the $MDTF$ is rotation invariant in planes in parallel to the chest wall using MRI exams with segmentations of fibroglandular tissue. For 424 breasts the mean $MDTF$ computed in 36 directions and the coefficient of variation were determined. As expected, the coefficient of variation decreases with an increase of $MDTF$. For extremely dense breasts, the coefficient of variation becomes as low as 0.05. This indicates that the method we propose is a feasible alternative to methods relying on a fatty tissue reference region when breasts are dense. It should also be noted that random errors due to variation of the direction in which the dense tissue fraction is measured are reduced because in mammography we combine four measurements, the MLO and CC projections of both breasts, into a single density measurement. For non-dense breasts the method cannot be used because the assumption of rotation invariance is violated.

A second assumption was made that the $MDTF$ in the direction of the compression is comparable to the $MDTF$ in a direction perpendicular to the compression force. Results in Figure 4.9B suggest that this is true in good approximation. For a moderately compressed breast, as obtained in MRI guided biopsy procedures, the $MDTF$ in both directions are comparable to each other.

Breast MRI scans were used for validation in which fibroglandular tissue, fatty tissue, and background were segmented. While some of these segmentations were obtained manually, most were automatically generated. That we used a mix of two segmentation methods may be regarded as a drawback since it could introduce an additional source of error. However, we observed that the range of volumetric breast density estimates of the manual segmented MRI's is comparable to the range of estimations obtained with the automatic segmented MR images (1.7-51.2% compared to 2.4-56.6%). Therefore, we believe that the manually segmented cases were not systematically denser of less dense than the automatically segmented cases.

In this study, we also investigated two different breast volume computation methods. Accurate estimation of breast volume is important because this is needed to obtain volumetric percent density. We found that a recently proposed method which extends the semi-circle

model of the breast edge shape leads to visually improved results when MRI is used as a reference. In this paper we adapted the method in such a way that it can be applied to both, MLO and CC, views. Both approaches resulted in a high correlation coefficient when compared to the breast volumes estimations computed from MRI data. However, the methods have a tendency to underestimate breast volume, but the bias is less prominent in the new method. This can be seen as well in Figure 4.8A where percent breast density is overestimated as the breast volume of the semi-circle model is underestimated. Some underestimation might be expected though, since the field of view of both imaging techniques is different. The imaged part of the breast in a mammogram relies strongly on the positioning of the patient. With MRI, the complete breasts and parts of the thorax are always imaged.

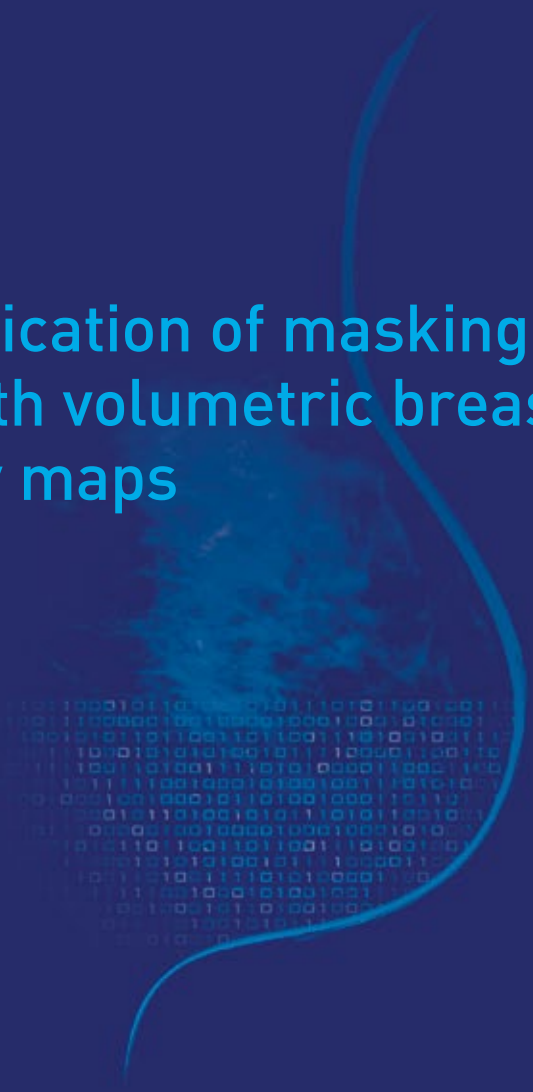
The algorithm of de Groot to estimate the breast volume defines a contact surface of the breast and the compression paddle that is slightly different from the one we used for defining the breast interior region for reference value $F2$. Though we could have used this, we decided not to use this alternative interior region definition. The reason is that we noted that due to inclusion of more pixels at the peripheral region border there was a high risk of including poorly compressed tissue regions close to the chest wall, especially in the MLO views. Inclusion of such regions might easily lead to large errors in the dense tissue estimation.

A limitation is that the study makes use of mammograms obtained at a single site with a single vendor. The studies of Wang [52] and Gubern-Mérida [53] have, however, shown that breast density can be reliable estimated with the reference method across vendors and that there is a problem with extremely dense breasts independent of the manufacturer of the mammographic system. Gubern-Mérida made use of images obtained with GE equipment, while Wang et al. used images acquired on Hologic systems. Hence, we do not expect problems with images from other vendors than GE.

In this study we proposed a novel approach to improve breast density estimation in dense breasts. By including the densest region in a mammogram as a reference and by adapting the region in which a fatty tissue reference region is selected we achieved better results compared to the existing approach. Results demonstrate that it remains crucial to find a good reference value for fatty tissue in order to get a reliable breast density estimate. Different fibroglandular tissue patterns within the breast require different techniques to estimate the fibroglandular tissue volume and percent volumetric density.

5

Quantification of masking risk with volumetric breast density maps



Original title: Quantification of masking risk in screening mammography with volumetric breast density maps.

K. Holland, C.H. van Gils, R.M. Mann and N. Karssemeijer

Published in: *Breast Cancer Research and Treatment*, 2017, 162:541-548

Abstract

Purpose: Fibroglandular tissue may mask breast cancers, thereby reducing the sensitivity of mammography. Here, we investigate methods for identification of women at high risk of a masked tumour, who could benefit from additional imaging.

Methods: The last negative screening mammograms of 111 women with interval cancer (IC) within 12 months after the examination and 1110 selected normal screening exams from women without cancer were used. From the mammograms, volumetric breast density maps were computed, which provide the dense tissue thickness for each pixel location. With these maps, three measurements were derived: 1) percent dense volume (PDV), 2) percent area where dense tissue thickness exceeds 1cm (PDA), 3) dense tissue masking model (DTMM). Breast density was scored by a breast radiologist using BI-RADS. Women with heterogeneously and extremely dense breasts were considered at high masking risk. For each masking measure, mammograms were divided into a high- and low-risk category, such that the same proportion of the controls is at high masking risk as with BI-RADS.

Results: Of the women with IC, 66.1%, 71.9%, 69.2% and 63.0% were categorised to be at high masking risk with PDV, PDA, DTMM, and BI-RADS, respectively, against 38.5% of the controls. The proportion of IC at high masking risk is statistically significantly different between BI-RADS and PDA (p-value 0.022). Differences between BI-RADS and PDV, or BI-RADS and DTMM, are not statistically significant.

Conclusion: Measures based on density maps, and in particular PDA, are promising tools to identify women at high risk for a masked cancer.

5.1 Introduction

Thanks to screening programs, breast cancers are often detected at an early stage. Nevertheless, not all breast cancers in breast cancer screening participants are actually detected by screening. Approximately 16-33% of the breast cancer cases are the so-called interval cancers, which means that they are diagnosed in between two screening rounds [14, 15], even though the introduction of digital mammography may have led to an increase in sensitivity [69, 106]. In general, interval cancers are detected at a later stage with a worse prognosis [16–18]. Fibroglandular tissue may mask cancers, and therefore sensitivity of mammography decreases with an increase in breast density. It has been shown that there is a relationship between breast density and screening program sensitivity [26, 29–32, 92]. In addition, compared to women in the lowest density category, women with dense breasts also have a higher breast cancer risk [10, 22, 24], which amplifies the negative effect of masking.

To detect more cancers at an early stage, personalised screening programs have been proposed [33, 34]. Adjusted to the personal needs of individual women, screening could be offered with different time intervals or with other modalities than mammography, such as ultrasound or MRI. Tomosynthesis might be an option as well, although the effect is limited for extremely dense breasts [107]. In this discussion, the reduced sensitivity of mammography due to the masking effect of density plays an important role. In recent years, many states in the United States passed breast density notifications laws. Radiologists are obliged to inform women about their breast density and the affiliated risks. In some states, additional imaging is reimbursed for women with dense breasts.

For the measurement of breast density, several methods are available. In clinical practice, the 4-point ACR BI-RADS scale is commonly used [37, 38]. To make this estimate less subjective, algorithms have been developed to estimate the breast density by computing the percentage dense area projected on the mammogram or by computing the percentage of fibroglandular tissue volume within the breast. An overview of different algorithms is presented by He et al. [54].

Although breast density relates to masking, the relation between the risk of masking and density is likely to be more complex than a simple dependence on the amount of fibroglandular tissue. Also, the distribution of dense tissue may play a role. This is reflected in the new BI-RADS definition that no longer considers the total amount of fibroglandular tissue within the breast, but rather the densest area [38]. How the risk of masking should be quantified is an open question. The aim of this study is to compare three different quantitative masking measurements and the visual BI-RADS density assessment of a radiologist, in their ability to predict the risk of an interval cancer.

5.2 Materials and Methods

5.2.1 Data

Digital mammograms from the Dutch breast cancer screening program were analysed. The mammograms were acquired on Lorad Selenia systems (Hologic, Bedford, USA). Women aged 50-75 years are invited biennially to participate in the screening program. Details about the screening program and the dataset can be found elsewhere [67, 108, 109]. Written informed consent was not required for this study. Women automatically consent to the use of their anonymised data for scientific purposes by participating in screening. Data of participants who objected to the use of their data were removed.

The research archive used contains unprocessed mammograms of one screening unit. In the period 2003-2012, more than 130,000 exams of more than 55,000 women were acquired. Mediolateral oblique (MLO) images were always taken, while craniocaudal (CC) images were taken in the first screening round and in 60% of subsequent rounds. Through linkage with the Netherlands Cancer Registry and the screening organisation, 1210 breast cancers were identified, of which 836 were screen-detected cancers. The remaining 374 breast cancers were diagnosed outside the screening program. Of these interval cancers, 275 were diagnosed within 24 months (screening interval), of which 113 cancers within 12 months after the examination. The last available screening examination before cancer diagnosis is used in this study. Women with breast implants were excluded from the study as the density map cannot be correctly computed for mammograms with implants.

In this study, a selection of the interval cancers, the cancers that were diagnosed within 12 months after the examination, is used (N=111, two women were excluded because of breast implants). The reason is that we want to focus on false negative exams due to masking. Interval cancers may also be due to other factors. In particular, fast growing cancers may not be detectable at the time of screening because they still are too small or not yet invasive. We assume that by excluding interval cancers detected more than 12 months after screening, a larger proportion of the interval cancers are due to masking. This idea is supported by Weber et al. [110] who found that a larger proportion of the interval cancers found in the second year after the screening examination show no signs of abnormality in the screening mammogram compared to the interval cancers found in the first year.

For each patient with an interval cancer, 10 participants were chosen as controls. The control participants needed to have had a mammographic examination in the same month in which the last screening examination of the interval cancer patient was performed. To be eligible as control, the women should not have been recalled on the basis of this mammographic examination and they should not have been diagnosed with breast cancer within 2 years after this examination. Women with breast implants were not eligible as control. Controls without a density map, due to failure of the computation, were replaced.

5.2.2 Methods

Quantitative masking risk measures based on volumetric breast density measurements were computed. For this purpose, a research version of the commercial software Volpara (v1.5.0, Volpara Health Technologies, Wellington, New Zealand) was used, which provides a quantitative breast density map in addition to the percentage of dense tissue volume. In these density maps, the pixel intensity is mapped to the fibroglandular tissue thickness at each pixel location.

Three different automated measurements were investigated to estimate masking risk: 1) percent dense volume (PDV), defined as the fibroglandular tissue volume divided by the breast volume; 2) percentage dense area (PDA), computed as the percentage area on the density map where the dense tissue thickness exceeded 1 cm; and 3) a dense tissue masking model (DTMM) in which the size distribution and cancer location probability are taken into account. The idea behind the second method is that a certain amount of fibroglandular tissue is necessary to hide a cancer. With the threshold of 1cm, the size of a region where cancers may be masked is estimated and we assume that the relative area of this region is related to masking risk. A strength of this method is the simplicity. In the third method, this idea is refined with the tissue masking model which captures two aspects. First, instead of using a fixed thickness threshold, it is modelled that larger cancers require more dense tissue to be masked than smaller cancers. For this, the normalised distribution of breast cancer size is taken into account. Second, the probability distribution of cancer location is used to consider that dense tissue presence in regions where cancers more often occur should give a stronger increase in masking risk than dense tissue presence elsewhere. A detailed description is in Box 5.1.

The methods were applied to all available images in an exam, i.e., MLO and CC views of both breasts. If CC views were missing, their results for the different methods were imputed. This was done for each method separately using linear regression analysis in controls with both MLO and CC views available. To come to a single score per exam, results were averaged over the four views.

Next to the automated measurements, for the purpose of this study, the breast density category of every exam was assessed by a radiologist (10 years of experience in breast imaging) according to the fifth edition of the BI-RADS atlas [38]. Mammograms were evaluated without knowledge of the cancer status.

To implement supplemental screening strategies in clinical or screening practice, it is necessary to divide the women into two groups: women at low masking risk and women at high masking risk. In practice, a threshold needs to be determined and all women with a measure above the threshold would receive additional imaging. The best threshold is unknown for the automated measures and depends on the screening population and the proportion of women that one is willing to offer supplemental screening or the number of interval cancers that should be detected with additional imaging.

Box 5.1 The dense tissue masking model

The dense tissue masking model (DTMM) captures two aspects: 1) the larger the lesion, the lower the masking risk; and 2) the larger the dense tissue thickness at a location, the higher the masking risk. Therefore, the masking risk for a lesion with diameter d_t at a location with density d is defined as

$$masking(d_t, d) = \frac{d - d_t}{d} \quad \text{if } d \geqslant d_t \tag{5.1}$$

From this, the masking risk for each pixel location (x,y) is estimated by summing over all possible tumour diameters considering the normalised cancer size distribution $s(d_t)$. The size distribution was obtained for invasive masses. In the mammograms, a contour was drawn around the mass and the effective diameter was determined (diameter = 2 (area/π)$^{1/2}$). The distribution was normalised to a value of one. With the use of the size distribution we take into account that lesions with the size of few millimetres are in general not detectable and that extremely large cancers are not common in screening. The masking risk at a location (x,y) with density $d(x,y)$ is then

$$\begin{aligned} masking(x,y) &= \sum_{d_t=0}^{d_t=d(x,y)} s(d_t) \times masking(d_t, d(x,y)) \\ &= \sum_{d_t=0}^{d_t=d(x,y)} s(d_t) \times \frac{d(x,y) - d_t}{d(x,y)} \end{aligned} \tag{5.2}$$

Furthermore, the cancer location probability distribution (CLPD) is taken into account. With the use of the CLPD, it is acknowledged that lesions are more common in the centre of the breast and less common close to the periphery. The CLPD was as well obtained with the invasive mass-like lesions as described in [111]. The probability of masking is then:

$$masking(x,y) = CLPD(x,y) \times \sum_{d_t=0}^{d_t=d(x,y)} s(d_t) \times \frac{d(x,y) - d_t}{d(x,y)} \tag{5.3}$$

With the formula above, the masking risk is determined for each pixel location (x,y). To come to a single score for each image, the masking risk is averaged over all pixels within the breast. Different CLPDs and size distributions were used for the MLO and CC images.

To measure to what extent the methods can identify women at high masking risk, the mammograms were divided in a high and low masking risk group by thresholding the risk measure. Then, the sensitivity of the masking measures was computed as the number of interval cancers in the high-risk group divided by the total number of interval cancers. The false positive rate is calculated as the percentage of normal controls selected as at high masking risk at the same threshold. In the context of risk stratification for supplemental screening, the proportion of controls selected as at high masking risk can be seen as supplemental screening rate and the proportion of interval cancers gives an estimate about the cancers that might be detectable with additional imaging at that supplemental screening rate.
The automated masking measures were compared to the radiologist scores when distinguishing BI-RADS density a and b versus BI-RADS density c and d. Bootstrapping was used to obtain 95% confidence intervals (CIs) and derive p-values.
Since breast density is a risk factor for breast cancer [10,22,24], it may be expected that the average breast density of women with cancer is higher than that in controls. Consequently, any predictive value of PDV for the presence of interval cancers might be caused by PDV being a risk factor, rather than being a 'masking factor'. To investigate the potential impact of this effect on our results, an additional experiment was conducted in which it was tested to what extent PDV can distinguish women with any breast cancer from controls. Again, cases with the highest PDV were selected by thresholding, and the proportion of cancers as a function of the proportion of controls selected was computed. For this experiment, mammograms of the screen-detected breast cancers and the interval cancers detected within 24 months were used. Only the interval cancers detected later than 24 months after the last examination were not used, as we assume that these cancers might well have been detected when women would have attended another screening round.

5.3 Results

The mean age of interval cancers and controls is 57.7 and 59.2 years, respectively. In 14.4% of the interval cancers, the cancer was diagnosed after first participation in the screening program, while 15.2% of the controls belong to women who attended the screening program for the first time. Only 3 interval cancers (2.7%) were diagnosed in women older than 70 years, while 110 women (11%) of the control group were older than 70 years.
In Figure 5.1, the percentage of interval cancers selected as at high masking risk is plotted against the percentage of controls selected when thresholding the different masking measures. As mentioned earlier, the proportion of controls selected as at high masking risk can be interpreted as the supplemental screening rate when a masking measure would be used in practice to identify women eligible for supplemental screening. The percentage of interval cancer selected as at high masking risk is a measure for the potential benefit of supplemental screening, since it is the proportion of women with interval cancers that would have

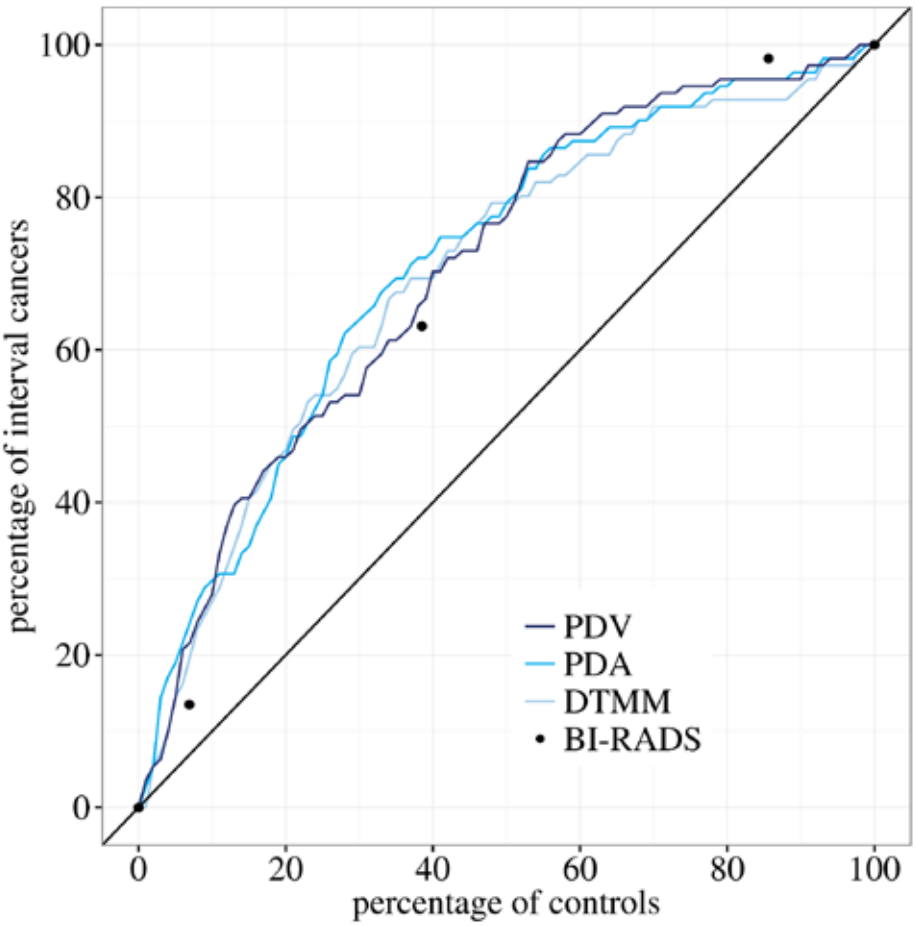


Figure 5.1: By thresholding the masking measures, cancers and controls were separated into high and low risk groups. The percentages of cancers and controls in the high-risk group are plotted against each other as function of the threshold.

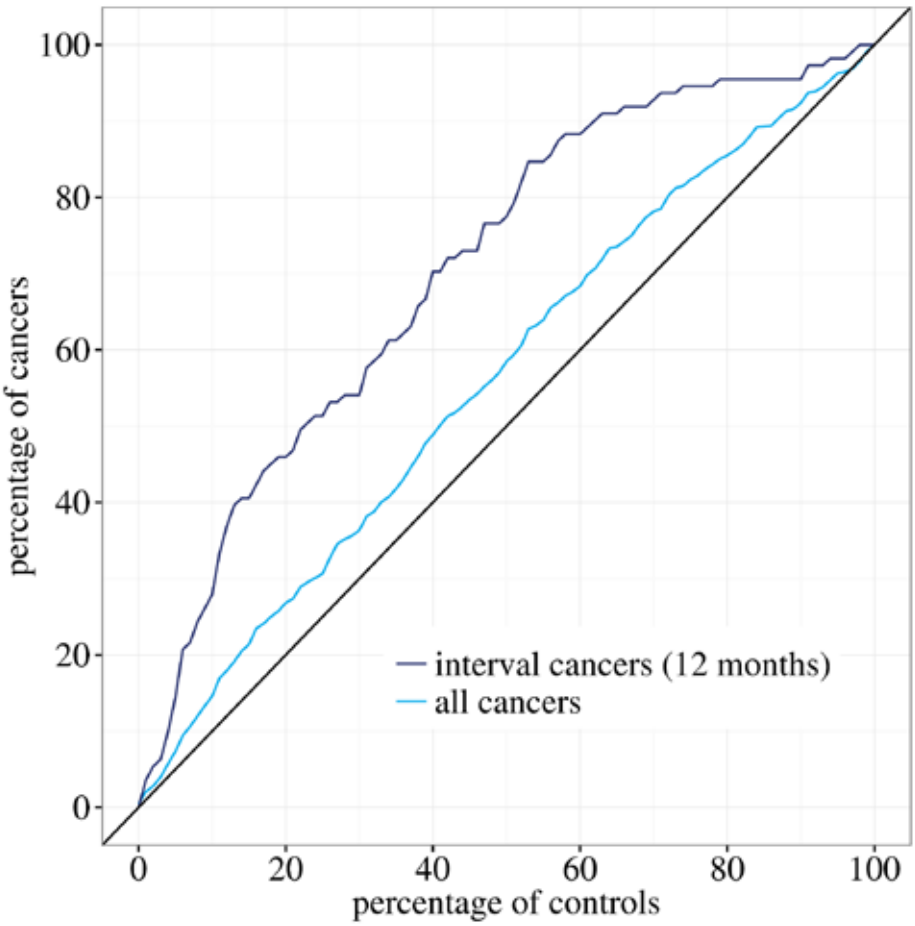


Figure 5.2: By thresholding the PDV measure the cancers and controls were separated into a high-risk and a low-risk group. The figure shows the proportion of cancers and controls in the high-risk group as function of the PDV threshold.

been included in the selection if supplemental screening had been offered. The percentage of interval cancers selected for several supplemental screening rates is given in Table 5.1A, while the supplemental screening rate required to include a certain percentage of women with interval cancer is presented in Table 5.1B.

The density scores determined visually by the radiologist were used to differentiate non-dense breasts from dense breasts, using the BI-RADS b-c transition as threshold. With BI-RADS, 38.5% (CI: 35.7-41.3) of the controls were considered dense and thus at high masking risk. Of the interval cancers, 63.0% (CI: 53.5-72.0) were classified as dense, using BI-RADS. If the thresholds of the three masking measurement methods were set such that there too the proportion of controls classified as at high masking risk was 38.5%, then 66.1% (CI: 55.8-76.2), 71.9% (CI: 63.1-80.2) and 69.2 (60.0-77.9) of the women with an interval cancer were considered at high masking risk with PDV, PDA and DTMM, respectively. Significantly more women with interval cancers would be included in the selection process with PDA compared to BI-RADS (p-value 0.022). Differences in proportions between BI-RADS and PDV, and BI-RADS and DTMM were not statistically significant with p-values of 0.187 and 0.067, respectively.

The ability of PDV to distinguish breast cancers from controls is displayed in Figure 5.2. The cancers detected at a screening examination (N=836) and the interval cancers that were diagnosed within the screening interval of 24 months after a negative screening examination (N=275) were eligible for the analysis (N=1,111). The PDV estimate was available for 1,103 cancers. The curve for predicting interval cancers shows a much higher area under the curve than the curve predicting all breast cancers. These results show that PDV is not 'just' a predictor for breast cancer risk, but in particular a good predictor for the risk of developing an interval cancer (as a proxy for risk of masking).

5.4 Discussion

In this study, we investigated the ability of several measurements of masking risk to distinguish false negative screening mammograms from true negative screening mammograms. The aim of our work is to find a method that is suited to identify women who are at high risk to be diagnosed with an interval cancer after a negative screening exam. In a personalised screening work flow, such a method could be applied to all negative screening mammograms to select the subgroup of women who would benefit most from additional imaging with MRI or ultrasound.

There are various reasons why interval cancers are not detected by screening, and masking is only one of them. Some cancers may be not detected by screening because they grow fast and the screening interval is too long. As we focus in this study on masking, we included in our experiment only those interval cancers that were diagnosed within 12 months after the negative mammogram, to exclude true interval cancers, the cancers that show no

A

percentage of controls	PDV	PDA	DTMM
5	14.4	18.9	14.4
10	27.9	29.7	27.0
15	40.5	34.2	40.5
20	45.9	45.9	47.7
30	54.1	64.0	60.4
38.5	66.1	71.9	69.2
40	70.3	73.0	69.4
50	77.5	79.3	79.3

B

percentage of interval cancers	PDV	PDA	DTMM
5	1.4	1.4	2.0
10	4.3	2.4	4.1
15	5.0	3.3	5.5
20	5.9	5.1	7.1
30	10.3	11.0	11.7
40	13.7	17.4	15.0
50	22.9	22.8	21.9
60	33.2	27.1	29.1
70	39.6	36.2	40.1
80	51.4	50.5	51.5
90	61.7	67.6	67.8

Table 5.1: On the masking measures, a threshold can be applied to separate cases and controls into a high-risk and a low- risk group. By adjusting the threshold on a masking measure the percentage of controls (also interpretable as supplemental screening rate) is adjusted. The percentage of interval cancers that would be included in the selection at several supplemental screening rates is given in subtable A. Using BI-RADS breast density c and d as high-risk categories, 38.5% of the controls are considered at increased masking risk and 63.0% of the women with interval cancer would be included in the selection. The threshold on the masking measure can be adjusted such that a specific percentage of the women with interval cancer is included in the high-risk group. The corresponding percentage of controls (supplemental screening rate) is given here for several percentages of interval cancers and the different masking measures in B. In total 111 cancers and 1110 controls were used.

signs of abnormality on the mammogram. True interval cancers are more common in the second year after the examination than in the first year [110]. Given that the exact cancer location was unknown and that the diagnostic mammograms were not available, it was not possible to review the interval cancers and to confirm that masking is the cause for a cancer diagnosis outside the screening program. It is noted that by excluding the interval cancers after 12 months, our study results are also more representative for screening programs with a 1-year interval.

In current clinical practice, the BI-RADS density assessment categories are used to decide whom to offer supplemental screening. Using a separation into a low-risk group (BI-RADS density a and b) and high-risk group (BI-RADS density c and d) with a 38.5% supplemental screening rate, it was found that between 63.0% and 71.9% of the women diagnosed with an interval cancer in the study data within 12 months of a negative screening would be included in the high-risk group for additional imaging. Automated measures have a higher sensitivity than the radiologist, and this difference was statistically significant for the new proposed measurement PDA at the chosen supplemental screening rate.

We compared the ability of PDV to distinguish cancers (screen-detected and interval) from controls to make sure that we capture more than the breast cancer risk in relation to breast density. Thereby, we confirmed that cancers are more common in dense breasts than in non-dense breast. Nevertheless, we can conclude that the differences in PDV distributions of interval cancers and controls are not only caused by the increased breast cancer risk that is associated with an increased breast density, and that PDV is capturing masking.

Cancers and controls were only matched for the month of acquisition and not for age and participation in the breast cancer screening program. The mean age of the controls was higher than the mean of the cases. Given that breast density decreases with age [25], one could argue that the difference in density distribution between cases and controls is caused by differences in age. However, the control group contained more women who participated in the screening program for the first time than the cases, leading to an effect in the opposite direction. While only three interval cancers (2.7%) were found in women between 71 and 75 years of age, 11% of the controls belong to this age group causing the higher mean age in controls. If we had matched for age at the time of acquisition, women above 70 years would have been under-represented in the controls, and the controls would have been not representative for the screening population.

Mainprize et al. [112] have been working on the quantification of masking as well. In their model, a detectability map is created for each pixel location by simulating lesions and by using local estimates of the noise power spectrum and volumetric breast density. They validated the masking measurement with an observer study on regions of interest of 150 cancer free CC mammograms. High correlations were found between the mean value of the detectability maps and the computerised and human observer study. However, Mainprize only used cancer free mammograms in his study and simulations in regions of interests. Hence,

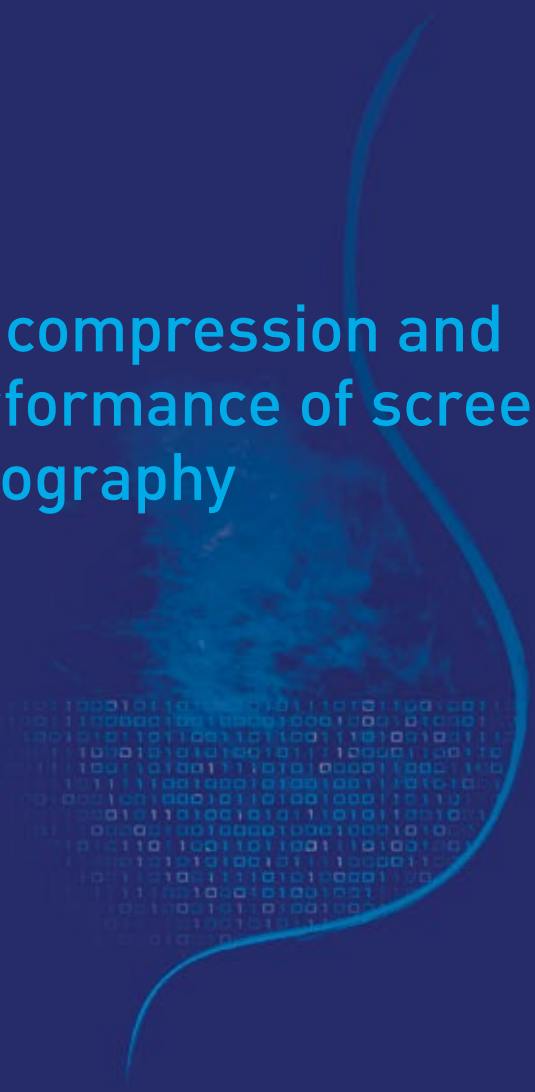
it remains an open question to which extent the mean value of the detectability map differs between false and true negative screening mammograms and whether it can be used as a predictive masking score.

A limitation of our study is the fact that CC images were not available for all exams. Until recently, MLO was the standard view in the screening program where we acquired our data, while CC images were obtained by indication. Therefore, to avoid bias when averaging over views, we imputed data for missing CC views based on the available MLO view and statistical analysis of differences between MLO and CC views. Furthermore, cases and controls were matched to the month of acquisition to guarantee the same guidelines and circumstance in image acquisition with regard to taking the CC views. Another limitation is that BI-RADS density assessments of only one radiologist were available. Many studies found inter- and intra-reader variability in breast density assessment using BI-RADS [41–43, 49]. Therefore, to make a definitive comparison between the automated methods and radiologists assessments, an extensive reader study should be conducted with multiple readers.

In conclusion, results suggest that the new proposed masking risk measurements may have a better performance than visual BI-RADS assessment in distinguishing false negative screening mammograms from true negative screening mammograms. Therefore, these measurements may be considered as predictive masking measure when implementing supplemental screening for women at a high risk for interval cancers.

6

Breast compression and the performance of screening mammography



Original title: Influence of breast compression pressure on the performance of population-based mammography screening.

K. Holland, I. Sechopoulos, R.M. Mann, G.J. den Heeten, C.H. van Gils and N. Karssemeijer

Submitted

Abstract

Purpose: To determine the effect of breast compression pressure in mammography on breast cancer screening outcomes.

Materials and Methods: We used digital image analysis methods to determine breast volume, percent dense volume and pressure from 132,776 examinations of 57,187 women participating in the Dutch population-based biennial breast cancer screening program. Pressure was estimated by dividing the compression force by the area of the contact surface between breast and compression paddle. The data was subdivided into quintiles of pressure and the number of screen-detected cancers, interval cancers, false positives, and true negatives were determined for each group. Generalized estimating equations were used to account for correlation between examinations of the same woman and for the effect of breast density and volume when estimating sensitivity, specificity, recall rate, false positive rate and positive predictive value. Sensitivity was computed using interval cancers occurring between two screening rounds and within 12 months after screening. Pair-wise testing for significant differences was performed for sensitivity and specificity measurements.

Results: Sensitivity in quintiles with increasing pressure was 82.0%, 77.1%, 79.8%, 71.1%, 70.8%. The 12 month sensitivity was significantly lower in the highest pressure quintile compared to the third (84.3% vs 93.9%, p = 0.034). Specificity was lower in the lowest pressure quintile ($p<0.005$) compared to the second, third and fourth group, with a specificity of 98.0%, 98.5%, 98.5%, 98.5%,and 98.4% respectively.

Conclusion: Results suggest that if too much pressure is applied during mammography this may reduce sensitivity. In contrast, if pressure is low this may decrease specificity.

6.1 Introduction

In mammography, breast compression is applied to reduce the thickness of the breast. This results in improved image quality, because tissue superposition and X-ray scatter are reduced, while it limits the required dose [113–116]. In addition, with a compression paddle the breast can be kept in a fixed position, which reduces the risk of motion artefacts and image blurring.

Mammography devices measure and display compression force during the imaging procedure. However, there are no quantitative guidelines regarding the compression force a radiographer should apply for acquisition of an adequate mammogram. In practice, compression force in mammography varies widely among radiographers, screening centre's, and countries [117–122]. A disadvantage of compression is that many women complain about discomfort and pain, which might influence their participation in screening [123–125]. A reduction in compression force has therefore been suggested to encourage screening attendance [126].

Although mammography systems display the compression force applied by the compression paddle to the breast, it is the pressure, which is defined as the compression force divided by the contact area between breast and compression paddle, that determines how much the tissue is compressed. It is therefore likely that the pain experienced by women undergoing mammography is more related to pressure than to force. The same force applied to a small or a large breast leads to different pressures. Pressure depends on the force and the contact area between the breast and the paddle, which depends on the breast size and the deformation and shape changes of the breast during compression. In case of a large contact area, the force is distributed over a larger area, leading to a lower pressure (force per unit area) compared to a small area. In a study by de Groot et al. [127] the force-standardised compression protocol was replaced by a pressure-standardised protocol using a recently developed paddle [128]. It was found that pain was reduced with pressure standardisation, while average glandular dose remained unchanged.

While it is widely accepted that firm breast compression is needed to ensure acceptable image quality, guidelines remain vague about how much compression should be applied [129] and a quantitative parameter indicating the amount of compression is not used. Consequently, little is known about the relationship between the amount of breast compression and breast cancer detectability. Furthermore, it has been reported that too much compression, as applied during spot compression, can lead to dissolving of suspicious densities in some cases [130–132]. Therefore, this retrospective study aims to investigate if the level of breast compression can impact screening performance, using pressure to characterise the level of compression.

6.2 Materials and Methods

6.2.1 Screening Data

In this retrospective study, mammograms that were acquired in the Dutch breast cancer screening program are used. In this population based program, women between 50 and 75 years of age are invited for a screening exam every two years. A consecutive series of mammograms acquired in one screening unit were collected. Raw mammograms acquired in this unit were archived between 2003-2012, except for a four month period in 2009 due to a technical issue. The images of the 135,640 available examinations were acquired on Lorad Selenia systems (Hologic, Inc., Danbury, CT, USA). Of these examinations, 113,956 examinations have been previously reported [133]. Cancer status information was obtained from the screening registration system and the Netherlands Cancer Registry. Written informed consent was not required for this study as women automatically consent to the use of their anonymised data for scientific purposes by participating in screening unless they object. Data of participants who objected to the use of their data was removed.

In this screening program, mediolateral oblique (MLO) and craniocaudal (CC) view images were always acquired at the first screening. However, during the study period, screening guidelines recommended that during subsequent rounds CC images be acquired only when the technician detected a possible abnormality or when high breast density was present. Therefore, CC view images are available only in about 60% of the exams. There are systematic differences between MLO and CC views in force and pressure. Therefore, to exclude these differences and to prevent a bias towards abnormal and dense breasts we used only MLO view images in this study.

Screen-detected cancers or true positives (TP) are defined as the cancers diagnosed after a recall of a woman for additional diagnostic tests. Interval cancers or false negative (FN) examinations are defined as cancers that were found within 24 months after a negative screening mammogram and before the next screening exam. False positive (FP) examinations are exams of women recalled for additional tests in which no breast cancer was diagnosed and true negatives (TN) are exams that did not lead to a recall and after which no breast cancer was diagnosed within 24 months before the attendance to another screening round.

6.2.2 Image Analysis

Compression pressure was determined retrospectively using a research version of the software Volpara (v1.5.0, Volpara Health Technologies, Wellington, New Zealand). The algorithm determines the contact area between the breast and the compression paddle by image analysis as described in [102]. The contact area measured with Volpara has been validated against manual segmentations from video images [134]. Pressure is computed by taking the compression force measurement from the imaging device, which is stored in the image header, and dividing it by the estimated contact area. Volpara software was also used to determine breast volume and dense tissue volume, which were used to compute percent

dense volume. For the true negative exams and bilateral findings, the values obtained for the left and right MLO views were averaged. For all other exams the values for the side with the finding or cancer was used.

6.2.3 Data analysis

The relationships between breast volume and force and pressure were investigated using heatmaps. The pressure measurements were used to divide the dataset into quintiles of increasing pressure. Within the five groups the following performance measures were determined: recall rate, false positive rate, screen-detected breast cancer rate, specificity, and positive predictive value as done in a previous publication [133]. Additionally, the interval cancer rate, the program sensitivity (number of screen-detected cancers divided by the sum of screen-detected cancers and interval cancers diagnosed before the next screening round) and 12 month sensitivity (program sensitivity using only the interval cancers diagnosed within 12 months after examination) were determined. The 12 month sensitivity may be used to translate results to the context of an annual screening program.
Many women in the data set had more than one screening examination. To account for correlation between examinations of the same woman, we used generalized estimating equations (GEE) using the 'independence' correlation structure. As it is known that the sensitivity of mammography is lower in women with higher breast density [29, 31, 92], breast volume and percent dense volume were included in the models to adjust for their potentially confounding effect. Breast volume and density were transformed using the natural logarithm to obtain data which approximated normal distributions. Pair-wise testing was applied to assess differences between the groups on the sensitivity and specificity measurements. Correction for multiple testing was applied using the Tukey method and a p-value below 0.05 was considered significant. GEE was used for each performance measure separately.
For each pressure group, the distribution of the different cancer types and their detection at screening or as interval cancers, using the non-corrected data, was also investigated. For this, the following categories were used: ductal carcinoma in situ (DCIS), invasive ductal carcinoma (IDC), invasive lobular carcinoma (ILC), 'other' and unknown.
Statistical analysis was performed using R version 3.3.1.

6.3 Results

In total 135,640 examinations were available. Excluding examinations with unknown screening outcome (N=72), examinations without a breast density, contact area or force measurement available (N=2,673), and interval cancers diagnosed more than 24 months after the examination (N=119), a total of 132,776 examinations were included in the analysis.
To stratify the exams into five groups of equal size, thresholds on the pressure estimates were applied at 7.7, 9.3, 10.8, and 13.0 kPa, resulting in the mean breast volumes, percent

breast densities, forces and pressures listed in Table 6.1. As expected, it can be seen that increasing compression pressure correlated with decreasing breast volume and increasing breast density, and force.

	Group 1	Group 2	Group 3	Group 4	Group 5
Number examinations	26,490	26,617	26,539	26,549	26,581
Mean breast volume (cm^3)	1511	1135	928	755	540
Mean percent dense volume (%)	5.7	6.6	7.5	8.4	10.7
Mean force (N)	112.6	121.5	125.9	130.7	138.9
Mean pressure (kPa)	6.6	8.5	10.0	11.8	15.6

Table 6.1: Number of screening examinations in each pressure group, and mean breast volumes, breast densities, forces and pressures in each group.

Heatmaps showing the variations in force and pressure with breast volume are shown in Figure 6.1, illustrating the difference between force and pressure. The horizontal lines in the pressure distribution mark the thresholds used to form the quintiles. It is observed that breasts of the same size are imaged using a wide range of forces. At the same time a trend is indicating that larger breast are imaged with higher forces, so some adjustment to the individual breast takes place. The pressure distribution shows, however, that the very large breasts are imaged with a low pressure and therefore most of these cases are in the first pressure category. Additionally, it is observed that the first pressure group contains the entire range of breast sizes, while extremely high compression is mainly a problem for small breasts. This is due to medium and high forces, as shown in the left heat map, being distributed over a small contact area, leading to high pressure.

Table 6.2 gives an overview of the screening outcomes for the complete cohort stratified by the five compression pressure groups. Screening performance measures are displayed in Table 6.3. Results suggest that at high compression pressure, sensitivity is reduced.

Results from the GEE models are shown in Table 6.4, confirming the decrease in sensitivity at high pressure observed in the unadjusted data. There is a statistically significant difference in the 12 month sensitivity between the third and the fifth group (p=0.0343). Even though this is the only significant difference between groups on the sensitivity measurements, a considerable difference can be observed between the first three pressure groups and the last two pressure groups. Results also show a trend that women with mammograms in the lowest pressure group are recalled more often. This leads to a higher false positive rate, and lower specificity and positive predictive value. The specificity was found to be significantly lower in the first group compared to the second (p=0.0002), third (p=0.0007) and fourth (p=0.0021) group. The 12 month sensitivity and the specificity are displayed in Figure 6.2 with and without correction for confounders.

In Figure 6.3, the distribution of DCIS, IDC, ILC and the remaining other types cancers

for the different pressure categories is given. The distribution is shown separately for all cancers, screen-detected cancers and interval cancers. As expected, only few DCIS cases are found among the interval cancers. The proportion of lobular and other types of cancers is relatively high for the interval cancers in the highest pressure group. Because of the low number of cancers in each subgroup, statistical analysis of subgroup differences was not performed.

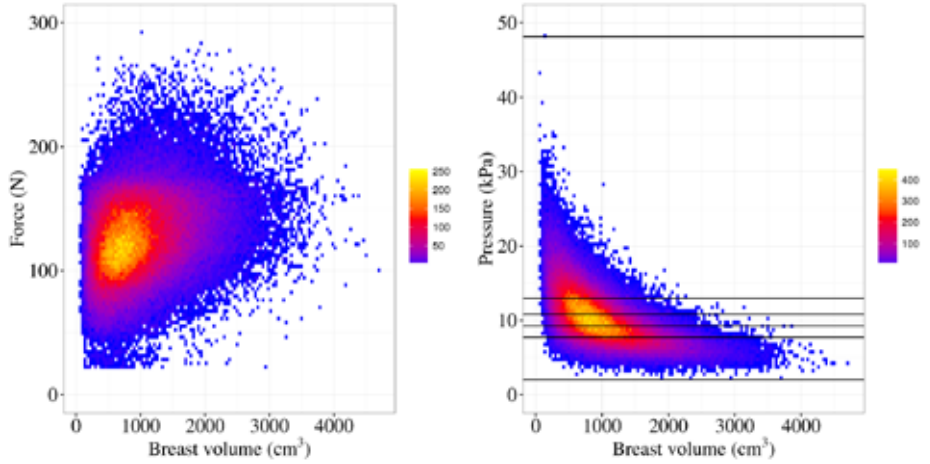


Figure 6.1: Measurements of force and pressure in relation to the breast volume. The colour code represents the number of exams in each bin. The horizontal lines in the right figure indicate the thresholds used to get the five pressure groups.

	Total	Group 1		Group 2		Group 3		Group 4		Group 5	
Total examinations	132,776	26,490	(20.0)	26,617	(20.0)	26,539	(20.0)	26,549	(20.0)	26,581	(20.0)
Screen detected cancers	833	177	(21.2)	164	(19.7)	188	(22.6)	152	(18.2)	152	(18.2)
Interval cancers	269	37	(13.8)	48	(17.8)	48	(17.8)	65	(24.2)	71	(26.4)
12 month interval cancers*	110	19	(17.3)	15	(13.6)	13	(11.8)	26	(23.6)	37	(33.6)
False positive examinations	2,192	486	(22.2)	381	(17.4)	404	(18.4)	426	(19.4)	495	(22.6)
True negative examinations	129,482	25,790	(19.9)	26,024	(20.1)	25,899	(20.0)	25,906	(20.0)	25,863	(20.0)

* Interval cancers detected within 12 month after the examination

Table 6.2: Number of screen-detected / interval cancers, false positive examinations and true negative examinations in the study population and in the five pressure classes. The 12 month interval cancers are included in the interval cancers. Within brackets is the row percentage.

	Total	Group 1	Group 2	Group 3	Group 4	Group 5
Recalls/1000	22.8	25.0	20.5	22.3	21.8	24.3
False positives/1000	16.5	18.4	14.3	15.2	16.1	18.6
Screen-detected cancers/1000	6.3	6.7	6.2	7.1	5.7	5.7
Interval cancers/1000	2.0	1.4	1.8	1.8	2.5	2.7
Program sensitivity (%)*	75.6	82.7	77.4	79.7	70.0	68.2
12 month sensitivity (%)**	88.3	90.3	91.6	93.5	85.4	80.4
Specificity (%)	98.3	98.2	98.6	98.5	98.4	98.1
Positive predictive value (%)	27.5	26.7	30.1	31.8	26.3	23.5

* Sensitivity calculated with all interval cancers

** Sensitivity calculated with the interval cancers that were detected within 12 months after the examination

Table 6.3: Unadjusted screening performance measurements.

	Group 1		Group 2		Group 3		Group 4		Group 5	
Recalls/1000	26.1	(24.0-28.5)	20.8	(19.1-22.6)	22.0	(20.3-23.9)	21.0	(19.3-22.8)	22.2	(20.3-24.3)
False positives/1000	20.0	(18.1-22.1)	14.8	(13.4-16.4)	15.1	(13.7-16.6)	15.2	(13.8-16.8)	16.2	(14.6-18.0)
Screen-detected cancers/1000	6.1	(5.2-7.2)	5.9	(5.0-6.9)	6.9	(6.0-8.0)	5.7	(4.9-6.7)	5.9	(4.9-7.0)
Interval cancers/1000	1.3	(0.9-1.8)	1.6	(1.2-2.2)	1.6	(1.2-2.1)	2.2	(1.7-2.8)	2.3	(1.7-3.0)
Program sensitivity (%)*	82.0	(75.6-87.0)	77.1	(70.9-82.4)	79.8	(74.2-84.4)	71.1	(64.5-79.8)	70.8	(63.6-77.1)
12 month sensitivity (%)**	90.1	(84.4-93.9)	92.0	(87.2-95.1)	93.9[a]	(89.7-96.5)	87.2	(81.3-91.4)	84.3[a]	(77.3-89.4)
Specificity (%)	98.0[a,b,c]	(97.8-98.2)	98.5[a]	(98.3-98.7)	98.5[b]	(98.3-98.6)	98.5[c]	(98.3-98.6)	98.4	(98.2-98.5)
Positive predictive value (%)	23.5	(20.2-27.3)	28.7	(25.0-32.7)	31.5	(27.9-35.3)	27.3	(23.7-31.1)	26.7	(22.9-30.8)

[a,b,c] Each subscript letter denotes a pair of pressure categories whose measurements differ significantly from each other

* Sensitivity calculated with all interval cancers

** Sensitivity calculated with the interval cancers that were detected within 12 months after the examination

Table 6.4: Screening performance measurements with 95% confidence intervals using GEE to account for correlations between mammograms of the same woman and to correct for the confounders percent dense volume and breast volume. Pair-wise testing was applied on the sensitivity and specificity measurements. The Tukey method was used for correction for multiple testing and a p-value below 0.05 was considered significant.

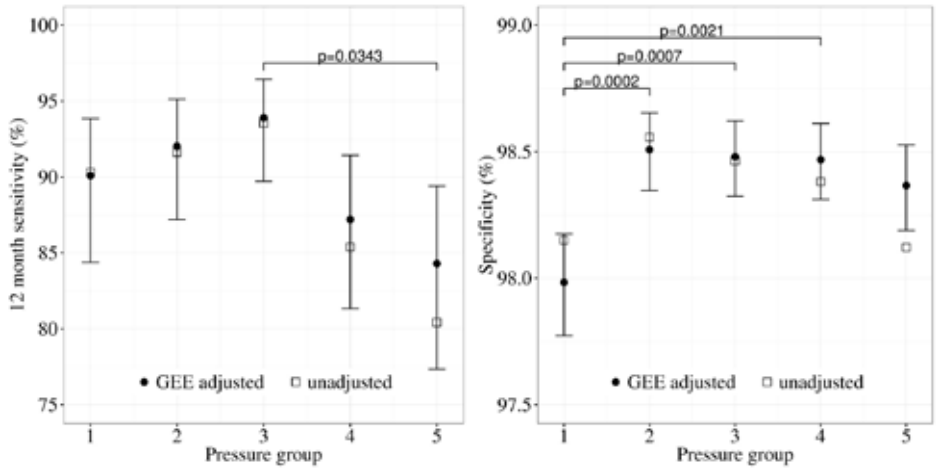


Figure 6.2: Measurements of 12 month sensitivity and specificity of the five pressure groups of unadjusted data (squares) and after adjustment with GEE for multiple screening rounds, breast volume and breast density including 95% CI (circles). Statistically significant differences between pairs of groups of the GEE adjusted data are indicated.

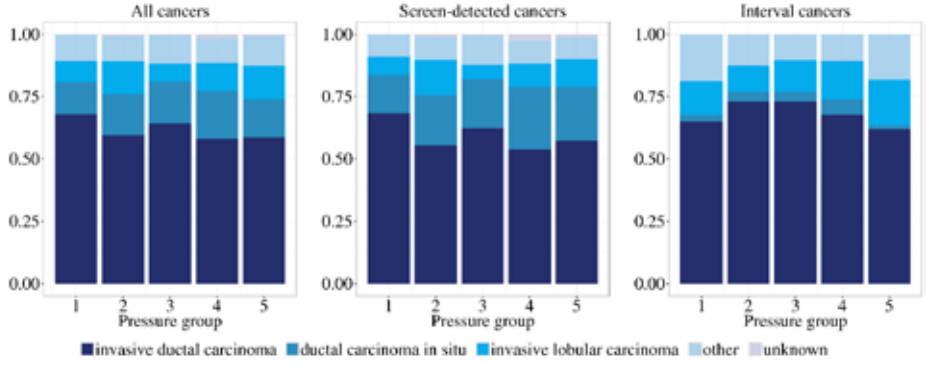


Figure 6.3: Cancer type distribution for the five pressure categories for all, screen-detected, and interval cancers. Because of the low number of cancers in each subgroup, statistical analysis of subgroup differences was not performed. The number of cancers for the five pressure groups, respectively, is 214, 212, 236, 217, and 223 for the all cancers together, 177, 164, 188, 152 and 152 for the screen-detected cancers and 37, 48, 48, 65 and 71 for the interval cancers.

6.4 Discussion

Given the reasons for using mechanical breast compression during mammographic imaging (reduction of tissue superposition, scatter, dose, possibility of motion, among other image quality-related benefits), it was expected that screening outcomes would be negatively impacted if the compression pressure applied was too low, as found in our results. As could have been foreseen, applying insufficient compression lowers the specificity of mammography, perhaps due to a lack of minimisation of tissue superposition. In this context, it should be highlighted that the overall compression force, and therefore pressure, is rather high in the Netherlands compared to other countries [117, 118, 120]. Hence, the loss of performance due to insufficient compression may be a more common issue in general than that found in this study. The finding demonstrates that an adequate level of compression is necessary to obtain good image quality and achieve a low recall rate, and stresses the need for techniques to apply compression at the right level. Though one should be careful with extending our conclusions to other populations, since screening policies vary from country to country, e.g. recall rates are higher in the US than in Europe, we note that our study sample is representative for breast cancer screening in Europe where low recall rates, as in our study, are common [135].

Though some reports were published about reduced visibility of a subset of tumours in spot compression [130–132], which typically are made with stronger compression, it was not a priori expected that high compression levels in screening mammography would have a large negative impact on screening outcomes. However, it seems that applying a higher than needed compression actually has a stronger negative effect on lesion visibility than applying insufficient compression, even when correcting for the confounding effects of breast density and volume, resulting in a lower sensitivity. Even though the difference in sensitivity did not reach statistical significance when corrected for confounding factors, except for the 12 month sensitivity between group 3 and group 5, the reduction was considerably larger than that suffered due to applying low compression.

It is not straightforward to identify the underlying cause for reduced sensitivity at high compression levels. This reduction in sensitivity could be due to either malignancies not being seen or being seen but mischaracterised, or due to both types of errors. Reduced visibility of certain tumours under high compression might be due to their composition. It can be reasoned that softer tumours may become less conspicuous with high compression, because the cancer tissue may spread out and lose contrast. Another reason could be that lesion types that are detected because of architectural changes of the breast parenchyma are less conspicuous under lower and higher pressure. This is supported by the distribution of ILC over the five groups. Although, it is hard to say with the given data whether a specific type of cancer is more often missed because of too low or too high pressure. The different cancer types should be kept in mind when investigating the relationship between pressure and screening performance in future studies, with other, perhaps larger, data sets. In terms

of mischaracterisation, vascularisation might play a role. Since invasive cancers are often highly vascularised, strong compression may lead to a reduction in blood flow [136, 137], leading to both a decrease in contrast and a reduction in the perceived suspiciousness of the finding, causing misinterpretation of its probability of malignancy [45].

The force and the pressure distributions were displayed in relation to the breast volume. The first pressure group contains the entire range of breast volumes. For women with small breasts, the low pressure is caused by a low force. For larger breasts, the contact area is larger so that even a medium or high force leads to a low pressure. An extremely high pressure is only observed for small breasts and is caused by too much force. Therefore, a compression recommendation based on force cannot solve the over-compression of small breasts and the under-compression of large breasts, as the force measure is independent of the individual breast characteristics. On the other hand, a pressure guided compression could prevent extreme compressions of small breasts and too low compressions of larger breasts, as the pressure measurement depends on the breast size, shape and stiffness.

One limitation of this study is that in MLO views, the pectoral muscle is included in the compressed tissue. Therefore, a portion of the compression force is absorbed by the muscle and not by the breast. Dustler et al. [138] found that the resulting pressure distribution during breast compression is not uniform, and that in some cases the pressure is highest in the pectoral muscle. Hence, estimating the pressure to the entire breast as a single value based on the overall force and area might not reflect the pressure sustained by the overall breast tissue, especially by the area where the missed lesions are located in the cases of reduced sensitivity. However, if the lesions were missed due to a decrease in blood flow into the breast tissue, perhaps the application of maximal pressure at the posterior border of the breast, against the muscle, is enough to have this effect.

Another limitation of this study is that CC views, and the per-view lesion sensitivity and specificity, were not included due to the fact that CC views were not acquired in a substantial portion of the screening exams at the time of data acquisition. In essence, this means that the presented results are based on the pressure applied for the acquisition of the MLO view only, while the screening outcomes are calculated based on the whole exam, which in the majority of exams also included CC views. To investigate the effect of compression in CC view images on the screening performance, a data set for which both views are available should be used. Finally, although this is a data set spanning many years and therefore includes exams in which various radiographers performed the acquisitions and various radiologists interpreted them, it is still a single-site study with all images acquired with a single mammographic system model. It is not expected, however, that compressions performed with other systems should have an impact on breast compression pressure and its relationship to screening outcomes.

Our retrospective analysis is performed with mammograms acquired in screening practice. Although the results and conclusions are based on a large sample of cases, evidence for our

findings would become stronger if mammograms of individual women could be obtained who were repeatedly imaged with different compressions. Since such evidence is currently lacking, we feel that it would not be appropriate to recommend an optimal pressure range based on this study alone. Further studies are needed to confirm our findings, ideally including a study with repeated imaging at different predefined pressures, to investigate lesion visibility as function of pressure.

In conclusion, this study shows a relation between the applied pressure and the performance of screening mammography even when taking into account confounding effects. The recall rate, false positive rate, and specificity were affected negatively in the compression category with the lowest pressure, while the sensitivity was reduced in the categories with high pressure. Since this is the first time this is reported, more research to confirm potential effects of pressure on performance is necessary because more attention to a meaningful standardisation of compression levels might improve mammography in the future.

7

Summary and discussion

Breast cancer is the most common cancer diagnosis in women [1]. Early detection is crucial to reduce breast cancer mortality. Therefore, many countries have implemented breast cancer screening programs, in which asymptomatic women of a certain age are screened regularly to find breast cancer before it becomes palpable and symptomatic. Most women are only screened with mammography during their life. Mammography has, however, its limitations. Fibroglandular tissue and cancer tissue have the same attenuation for X-rays making it hard to distinguish them on mammograms. Hence, it is possible that a cancer is masked in dense tissue structures and is therefore not detected with mammography. This leads to cancers that are detected outside the screening programs in between two screening rounds. These cancers, which are called interval cancers, are often diagnosed in a later stage with a worse prognosis [16–18]. To reduce the number of interval cancers screening should be improved. This may be accomplished by introducing personalised screening.
In personalised screening, the risk of development of breast cancer and the risk of missing a cancer with mammography is determined and subsequently used to offer individual women a screening procedure that best fits to their needs. Personalised screening can be implemented in different ways: by offering examinations at different time intervals, depending on the risk of developing breast cancer, by offering screening with different modalities, or by offering additional imaging to mammography in case of a decreased sensitivity with mammography.
This thesis focuses on mammographic breast density and its potential to be used as stratification tool for personalised breast cancer screening.

In Chapter 2, the performance of the Dutch breast cancer screening program is studied in relation to four breast density categories. We used the four volumetric density groups (VDG) obtained from mammograms by processing with the automated software method Volpara. It is observed that the risk of developing breast cancer increases with an increase in breast density (breast cancer rates of 4.6, 8.3, 9.4, 11.2 per 1000 women screened for the four groups respectively). At the same time, only a small increase is observed in the screen-detected cancer rate throughout the second to fourth density category (4.0, 6.4, 6.6 and 6.8 screen-detected cancers per 1000 women screened for the four groups, respectively), while strong a increase is observed in the number of interval cancers with increasing breast density (interval cancer rate of 0.7, 1.9, 2.9 and 4.4 per 1000 examinations). Together this leads to a substantial drop in screening program sensitivity (85.7%, 77.6%, 69.5%, 61.0%) when breast density increases, despite the higher recall rates observed with increasing breast density. Thus, women with dense breasts do not only have an increased breast cancer risk but also an increased risk that the cancer is not detected with mammography within the national breast cancer screening program.
In the past years, mammographic breast density has become a major issue in breast screening practices the United States. Many states passed legislation that requires radiologists

to inform women about their breast density and the associated risk for breast cancer and masking. In this context, the ACR BI-RADS density categories 1 and 2 (4th edition) or a and b (5th edition) are considered non-dense, while women with a BI-RADS density category 3 or 4 (4th edition) or c or d (5th edition) are informed that they have dense breasts. Additional imaging with ultrasound in screening exams of women with dense breasts is reimbursed in some states [139]. To use such a categorisation into four or two categories to personalise screening, it is necessary to have a consistent categorisation, also over time. Otherwise, women and clinicians might lose confidence in the stratification process. This is acknowledged in a review paper, where concerns are raised that radiologists' variability of BI-RADS density assessments over time may lead to inconsistent information in mandated communications about elevated breast cancer risk and supplemental screening [62]. Several studies show that the BI-RADS categories are prone to inter-reader and intra-reader variability [39–42]. The use of automated software would solve the problem of intra-reader variability as the algorithm always gives the same result.

In the study described in Chapter 3, breast density assessment of serial mammogram pairs is investigated and the results of four readers are compared to the automated density assessment with Volpara. First of all, it was found that when using a binary classification into non-dense and dense breasts, most pairs get assigned to the same category. The expected decrease of breast density with age is observed in this study as well, about 6-10% of the women switch density categories from dense to non-dense in subsequent screenings. A change in category from non-dense to dense is not expected frequently because this does not fit with the regular biological pattern of changes in the breast. With software, less category changes into the dense category are observed compared to the human assessment. In screening practice, consecutive mammograms are usually not read by the same radiologist. Therefore, to measure consistency over time of human assessment of breast density, in the analysis of our reader study data, each exam was randomly assigned to one of the readers to simulate screening practice. It was found that the software has significantly less category changes and a higher agreement between the two mammograms than human readers. Therefore, software based assessment may be preferred when estimating breast density for breast density stratified screening.

With the introduction of digital mammography, it became possible to save more information than the X-ray image. Nowadays, compression settings, and tube and filter material information are stored together with the pixel data. Using an internal calibration and raw image data it is possible to accurately determine the volume of fibroglandular tissue in the breast [45, 46]. This method is implemented in the Volpara software. Thus, from the 2-dimensional mammographic image a 3-dimensional quantity is obtained. To evaluate the performance of volumetric breast density assessment algorithms, mammographic breast density is compared to breast density measured with MRI. These comparisons show, that with the internal calibration, breast density is underestimated in extremely

dense breasts [51–53]. This is caused by the fact that a pixel with the least amount of fibroglandular tissue is used as a reference. When the breast is non-dense, it can be assumed that this pixel belongs to the projection of only fatty tissue, but in an extremely dense breast, this assumption is not valid anymore. As a consequence, the fibroglandular tissue volume and subsequently percent density are underestimated. In Chapter 4, a novel method is presented to improve breast density estimates for dense breasts. In case the breast is non-dense, the standard method is used, that has been validated many times. The new method assumes that the breast is very dense. For this reason, the new method can not be applied to non-dense breasts and it is necessary to use a combination of methods to obtain a breast density estimate for all types of breasts. The best results were obtained with the combination of methods. It is noted that the internal calibration approach used in our work is not the only method to obtain a volumetric breast density estimate for the mammogram. But, compared to other methods our approach has the advantage that only a mammogram with the 'raw' pixel data is required to apply the method. Other methods require that a phantom is placed next to the breast during mammogram acquisition or a complex calibration of the imaging unit [78, 79, 99].

Chapter 2 shows that the sensitivity of mammography decreases with an increase in breast density and that more interval cancers were found at higher densities than at lower densities. This effect is explained by masking. Fibroglandular tissue and cancerous tissue have the same attenuation for X-rays. Hence, both types of tissue have the same appearance on the mammogram. In case of a dense breast, it is then more difficult or not possible to detect the cancer in the mammogram. The cancer is obscured by dense tissue structures and masks its true nature. The explanation that interval cancers are also found in non-dense breasts is given by the definition of interval cancers. Interval cancers are defined as the cancers that are found after a negative screening round and before the next scheduled examination. This definition does not take into account the presence of the cancer at the last examination. Therefore, it is also possible that the cancer is a fast growing lesion that was not present in the breast at the time of screening and was therefore not detectable by the radiologist. These cancers are called true interval cancers. To decide whether the interval cancer is a true interval cancer or not, it is necessary to compare the diagnostic mammogram with the last screening mammogram. Studies have shown that many interval cancers are visible on the screening mammogram in retrospect [19–21]. The problem is, however, that not all cancers that are visible on the screening mammogram in retrospect are masked cancers. It is also possible that the lesion was simply missed by the radiologist. Furthermore, it is also possible that the lesion is already present in the breast and that there are no signs of malignancy in the mammogram, as it is completely impossible to distinguish cancerous from fibroglandular tissue.

Although the presence of breast density is related to masking, the relation is likely to be more complex than a simple dependence on the amount of density. Also, the distribution

of dense tissue may play a role, which is reflected in the fifth BI-RADS definition. In Chapter 5, it is investigated to which extent volumetric breast density can be used to predict an interval cancer as a proxy for masking risk. This was done with the interval cancers that are detected within 12 months after the negative screening mammogram, to increase the likelihood that the lesion was already present in the breast at the time of screening. It was a limitation of our study that the diagnostic mammograms were not available to classify all interval cancers as true interval cancers or as undetected cancers. Next to percent dense volume, two other automated measurements are investigated to predict interval cancers. These measurements make use of density maps. A density map represents the amount of fibroglandular tissue for each pixel location. The study shows that a simple measurement such as the percentage breast area that is covered with a fibroglandular tissue column of more than 1cm might be a better masking risk predictor than percent dense volume. Moreover, a dense tissue masking model is presented. The model takes into account the distribution of tumour diameters of screen detected cancers and a cancer location probability map. With the model, it is acknowledged that the risk of masking increases with increasing (local) breast density while it takes into account that the risk of masking decreases with increasing cancer size. The cancer location probability distribution considers that a cancer is more likely to be present in the interior region than in the breast periphery. Unfortunately, we found that though the masking model is more sophisticated than thresholding of the density map, the performance remains behind the thresholding method. All automated methods show a better performance in predicting interval cancers than an experienced radiologist did with thresholding at the BI-RADS b/c transition.

The breast is compressed between the compression paddle and the detector during mammogram acquisition to reduce dose, tissue superposition, X-ray scatter and the possibility of motion and therefore motion artefacts [113–116]. Today, compression is often force controlled. The force applied is independent of breast size and many women, especially those with small breasts, complain about pain during the acquisition. A better measurement of compression might be pressure, defined as force divided by contact area. The contact area is the area of the breast that is in contact with the compression paddle. In Chapter 6, the screening program performance is evaluated in relation to pressure. The data is divided into five groups of equal size depending on the pressure estimate of the MLO view with a finding. Generalized estimating equations are used to determine the screening performance measures and to account for differences in breast volume and breast density within the groups. It is found that the women in the first pressure category are more often recalled. In this group also a higher false positive rate and a lower positive predictive value was observed. These results support the necessity of sufficient compression to guarantee image quality and avoid unnecessary recalls. On the other hand, it is found that too much compression is contra productive for the screening program performance. The sensitivity is lowest in the two groups with highest pressure, e.g. more interval cancers are found in

these groups. Based on these results, which should be confirmed by other studies, one could consider to change recommendations in screening. Instead of applying approximately the same force to all breasts, compression should be adjusted to the individual breast volume and composition with a pressure measurement.

Breast density: Are we ready for breast density stratified screening?

In the previous chapters, we have looked at different aspects of breast density, how to measure breast density and how it would perform when used for stratification. In this last section, we will discuss whether personalised screening is feasible within the next years in Europe and what needs to be done to implement breast density stratified screening.

The idea to personalise breast cancer screening based on breast density arises from the fact that women with dense breasts have a higher masking risk and a higher breast cancer risk compared to women with non-dense breasts as seen in Chapter 2.

The first problem that comes up when introducing personalised screening is the missing agreed definition of breast density that is suited for high volume stratification. Different breast density measurements were used in the past and they all show the increase in breast cancer risk with increasing breast density. So far it remains unclear which breast density measurement should be used for risk assessment. In clinical practice, the BI-RADS categories are commonly used. The problem with BI-RADS categories is the large inter- and intra-reader variability. The use of automated software is a way to remove intra-reader variability and to make personalised screening procedures objective and predictable. To eliminate 'inter-reader' variability it is necessary to choose one algorithm.

For a long time, the semi-automated software Cumulus [55] was considered the gold standard and a strong relationship between Cumulus density measurements and breast cancer risk were observed. The problem with Cumulus is, however, the fact that a percentage dense area is measured, which is not projection invariant, and that Cumulus is semi-automatic, so it requires still some user input which makes it labour intensive and prone to intra- and inter-reader variabilities. Hence, Cumulus is less suited for screening where fast, consistent and objective measurements of breast density are needed. Therefore, automated volumetric measurements are a better option. But, even when deciding to use volumetric breast density, a wide palette of options and algorithms is available. Starting by algorithms that use an internal calibration [45–47], like the commercially available programs Volpara and Quantra, and continuing with algorithms that require a phantom to be placed next to the breast [99] or a complex calibration of the mammographic system [78, 79]. Even more options are described in the review of He et al. [54].

When a method for breast density measurement has been chosen, it still remains open how to use these measurements for risk stratification. Is percentage dense volume the best measure, or should the fibroglandular tissue volume and breast volume be combined in a different way? Should we estimate the risk with continuous measurements, should we use quartiles,

or should we use some other thresholds, like the ones used to get the VDG categories out of the percent density measurement? Should the density and risk be estimated based on images of the first screening round or with each screening round?

To answer these questions researchers are actively looking into these issues. In the study by Eng et al. [24], six breast density algorithms were used to estimate the breast cancer risk. Commercially available percent dense volume estimations were compared to BI-RADS and Cumulus. It was found that the highest risk predictions were found with automated measurements of volumetric percent density, in particular with Volpara. In the study by Wanders et al. [140], different combinations of Volpara results were used to estimate the breast cancer risk (using all cancers within the database) and masking risk (using the interval cancers). It was found that the fibroglandular tissue volume gave higher risk estimations than percent dense volume when estimating the breast cancer risk, while risk predictions for interval cancers were highest when using percent dense volume. Therefore, the density measurement used for stratification could be different for the different risk estimations. As personalised screening aims for a reduction of the interval cancer rate and for finding cancers that are otherwise missed with mammography, percent dense volume could be preferred for risk estimations and stratification.

A difference between the studies of Eng and Wanders is the evaluated examination. Eng et al. made use of the last available mammogram before the cancer diagnosis, while Wanders et al. used the first digital available mammogram. Given these results, it is not clear on which measurement risk stratification should be based. When using the first available mammogram, the breast cancer risk is only estimated once. In that case, the entire screening procedure is determined based on one estimate. As the density decreases with time, also the risk of masking decreases as there is less fibroglandular tissue to obscure the lesion. Therefore, taking the most recent mammogram to determine masking risk is a logical choice as the screening procedure should be adjusted to the risk at the time of screening. This way of implementation has also its drawbacks. Variability over time is possible, which could lead to confidence loss in the stratification process. Our study in Chapter 3 is of relevance in this respect, as it shows that serial mammograms are assigned to the same density class more often by the Volpara software than by human readers.

In the study by Eng et al., only the four point categorisation VDG was used, while Wanders et al. presented results for VDG and the continuous percent dense volume measurement. The underestimation of percent density in extremely dense breasts, that was discussed in Chapter 4, is not a problem when using VDG, as the thresholds to form the groups are chosen in such a way that the breasts with underestimation are still in the highest density category. However, when using continuous measurements like in Wanders et al., the risk estimation might benefit from the improvement in fibroglandular tissue volume estimation as obtained with the novel method proposed in Chapter 4.

Though breast density shows a higher correlation with breast cancer risk and masking risk,

it is likely that the risk of masking is more complex and that a measurement of percent dense volume is not enough. We showed that other measurements perform better in the differentiation of interval cancers from true negative screening mammograms and that these might be preferably for stratification in a personalised screening program (Chapter 5). Furthermore, combining texture measurements with breast density estimations can improve risk estimations [141].

The well established fact of decreasing sensitivity of screening with increasing breast density has already led to breast density legislation that mandates the disclosure of breast density to women undergoing mammography screening in the United States. In case of dense breasts, increased breast cancer risk and masking risk as well as supplemental screening options are communicated with the client. In some states, additional imaging is reimbursed. But, laws vary from state to state leading to different implications for the women [142]. Results of the changing screening practice in the U.S. are starting to appear in the scientific literature and it is likely that these will have an impact on screening policies elsewhere. Also in Europe, the use of alternative screening modalities such as breast ultrasound, breast tomosynthesis, and breast MRI is increasing.

A current problem is that the evidence for cost-effectiveness and benefits of breast density stratified screening is not that clear. The review of Melnikow et al. [62] states a need for well-designed, long-term and prospective studies. There are studies that investigate the benefit of ultrasound or MRI after a negative mammogram [58–60]. The problem with these studies is that only the cancer rate and the false positive rate with and without intervention are measured. Additional imaging increases the cancer rate at the cost of an increase in the false positive rate. It remains open how breast cancer mortality and interval cancer rate are affected. A step in the right direction is the DENSE trial [61], which is ongoing in the Dutch breast cancer screening program. In the randomised controlled trial, MRI is offered to women with extremely dense breasts (VDG 4). The primary outcome measure is the difference in proportion of interval cancers. First results of this trial are expected within the next years.

To conclude, to introduce breast density stratified screening in Europe, more evidence is needed that additional imaging improves the patient outcome with low additional burden for the screening population at low costs. As interval cancers are usually detected in a later stage, a reduction of interval cancers could be used as intermediate measure. Furthermore, the research community has to agree on acceptable ways to measure breast density, breast cancer risk and masking risks. Last, we have to think carefully about patient communication. It should be clear how stratification is performed and which implications it has. In the end, personalised screening should improve the current screening regime. The benefits should outbalance the harms.

Samenvatting

Borstkanker is de meest voorkomende kanker bij vrouwen en de vroege opsporing van borstkanker in het bevolkingsonderzoek verhoogt de kans op overleving. In dit bevolkingsonderzoek krijgen asymptomatische vrouwen wanneer ze in een bepaalde leeftijdcategorie zijn regelmatig een mammografisch onderzoek aangeboden. In de meeste landen begint screening rond de leeftijd van 50 jaar, en de laatste uitnodiging om aan het bevolkingsonderzoek deel te nemen wordt ontvangen door vrouwen met een leeftijd van 70 of 75 jaar. In deze leeftijdgroep is de kans op het ontwikkelen van borstkanker het hoogst. In Nederland krijgen vrouwen iedere twee jaar een mammogram aangeboden wanneer ze tussen 50 en 75 jaar oud zijn.

Vele studies hebben de effectiviteit van borstkankerscreening aangetoond, maar er zijn niet alleen voordelen. Naast het vroegtijdig opsporen van kanker bestaat er ook de kans op vals alarm, zogenaamde fout-positieve uitslagen. Daarnaast betekent een negatieve uitslag niet dat er geen kanker aanwezig is in de borst. Ruim 16-33% van de borstkankers wordt buiten het bevolkingsonderzoek tussen twee screeningsrondes gediagnosticeerd. Deze kankers worden intervalkankers genoemd en zijn in sommige gevallen al zichtbaar op het voorgaande mammogram.

In het huidige programma krijgen alle vrouwen hetzelfde onderzoek aangeboden: een mammogram. Het risico op het ontwikkelen van borstkanker is echter niet voor alle vrouwen hetzelfde. Het borstkankerrisico hangt af van vele factoren. De verschillende risicofactoren zouden gebruikt kunnen worden voor een individueel risicoprofiel dat vervolgens voor een 'screening op maat' gebruikt kan worden. Afhankelijk van het risico op borstkanker en het risico dat de kanker met een bepaalde techniek gemist wordt, kan dan de screening geïndividualiseerd worden door verschillende vrouwen verschillende screeningstechnieken met verschillende periodes tussen de screeningsrondes aan te bieden.

Een van de borstkankerrisicofactoren is borstdichtheid. De borstdichtheid is een maat voor de hoeveelheid klierweefsel in de borst. Tegenwoordig, in de tijd van de digitale mammografie, is het mogelijk het klierweefselvolume te berekenen door een computer. Samen met een berekening van het borstvolume kan dan het percentage klierweefsel geschat worden. In de klinische praktijk wordt meestal gebruik gemaakt van een dichtheidsschatting van een radioloog. Met behulp van de ACR BI-RADS atlas bepaalt de radioloog tot welke van de vier mogelijke dichtheidscategorieën het mammogram hoort. Het probleem met de schattingen van radiologen is echter dat de categorieën verschillen wanneer verschillende radiologen hetzelfde mammogram hebben beoordeeld, en dat ook de categorieën kunnen verschillen wanneer dezelfde persoon een beeld op verschillende tijdstippen heeft beoordeeld. Door de dichtheidsschatting door een computer te gebruiken is het mogelijk de eerder genoemde inter- en intralezervariabiliteit te voorkomen. Om de automatische schatting met de BI-RADS categorieën te kunnen vergelijken, is het nodig om van het klierweefselpercentage over te gaan naar een categorische maat. Dit klierweefselpercentage onderverdeeld in vier klassen wordt uitgedrukt als 'Volumetric Density Grade' (VDG). VDG is

een dichtheidsschaal die gebaseerd is op het percentage klierweefsel gemeten met Volpara, een commercieel beschikbaar softwarepakket. Studies hebben laten zien dat de dichtheidsschatting van een radioloog met BI-RADS en VDG goed correleren.

Een maat om de prestaties van borstkankerdetectie in het bevolkingsonderzoek te meten is sensitiviteit; het percentage borstkankers dat binnen screening wordt ontdekt. In hoofdstuk 2 lieten we zien dat de sensitiviteit afhangt van de borstdichtheid. Hoe hoger het klierweefselpercentage, des te groter de kans dat een tumor buiten het bevolkingsonderzoek wordt ontdekt. Vrouwen met een hoge borstdichtheid hebben dus naast een verhoogd risico op het ontwikkelen van borstkanker ook een verhoogd risico op het niet ontdekken van de kanker binnen het bevolkingsonderzoek, in vergelijking met vrouwen met een lage borstdichtheid. Ook werden vrouwen met veel klierweefsel vaker doorverwezen en hebben vaker een fout-positieve uitslag.
De discussie over lage sensitiviteit ten gevolge van hoge borstdichtheid heeft in de Verenigde Staten (VS) tot wetgeving geleid, de zogenoemde 'breast density notification laws'. In veel staten in de VS worden vrouwen over hun borstdichtheid en de daarmee verbonden risico's geïnformeerd en in sommige staten krijgen vrouwen additionele beeldgeving aangeboden. Vaak wordt er alleen onderscheid gemaakt tussen een lage en hoge borstdichtheid (BI-RADS a/b versus c/d, VDG 1/2 versus 3/4). Op deze manier kan borstkankerscreening individueel worden afgestemd. Afhankelijk van het individuele borstkankerrisico en het risico op een intervalkanker kan de mammografie door een andere modaliteit vervangen worden of wordt er additionele beeldvorming aangeboden zoals tomosynthese, echografie of MRI.
In hoofdstuk 3 hebben we gekeken hoe de borstdichtheidsschatting zich gedraagt naarmate de leeftijd hoger wordt. Dit is van belang wanneer de borstdichtheid wordt gebruikt om te stratificeren naar individuele screeningsschema's, omdat afwijkende adviezen tussen de tweejaarlijkse screeningsronden mogelijk kunnen leiden tot onzekerheid en een verminderd vertrouwen in de stratificatie. Om hier inzicht te geven in de eventueel veranderende borstdichtheid van vrouwen hebben we van 500 vrouwen borstdichtheid van twee opeenvolgende screening mammogrammen geschat. Dit werd zowel met geautomatiseerd software gedaan (VDG) als ook door radiologen die de BI-RADS classificatie gebruikten. We hebben met dit onderzoek laten zien, dat in meer dan 85% van de gevallen opeenvolgende screeningsmammogrammen in dezelfde borstdichtheidscategorie terecht kwamen (laag versus hoog) wanneer één radioloog beide screeningsrondes beoordeelde. Maar in de praktijk worden binnen het bevolkingsonderzoek verschillende screeningsrondes vaak door verschillende radiologen beoordeeld. Om die reden hebben we een situatie gesimuleerd waarin de borstdichtheid van iedere screeningsronde van een willekeurige radioloog visueel geschat werd. Het blijkt dat in deze situatie twee opeenvolgende screeningsrondes vaker in dezelfde dichtheidsgroep terecht komen met de door software automatisch bepaalde VDG groepen dan wanneer ra-

diologen visueel de BI-RADS classificaties gebruiken. Daaruit concluderen we dat het misschien beter is om de dichtheidsschattingen met software te doen dan in plaats van een radioloog. Het voordeel van software is bovendien dat het resultaat altijd hetzelfde is (tenminste als het dezelfde algoritme gebruikt wordt), er is dus geen intralezervariabiliteit.
Softwarepakketten zoals Volpara bepalen zowel het borstklierweefselvolume als ook het borstvolume op basis van een het mammogram, de instellingen van de apparatuur, bijvoorbeeld het materiaal van de röntgenbuis en de elektrische spanning, en de dikte van de gecomprimeerde borst. Ter verificatie van deze schatting, wordt de mammografische dichtheid vaak vergelijken met de dichtheidsschatting uit MRI beelden. MRI is een beeldvormende techniek welke driedimensionale (3D) beelden van het lichaam kan maken, waardoor het makkelijker en nauwkeuriger is om een volume te berekenen. Voor dit doel worden voxels (3D versie van een pixel) met klierweefsel van voxels met vetweefsel onderscheiden. Om het klierweefselvolume de berekenen gebruikt het Volpara algoritme voor elke mammografische opname een aparte referentiewaarde. Daarvoor gaat het algoritme op zoek naar de donkerste pixels in het mammogram. Het algoritme veronderstelt vervolgens dat deze pixels bij de projectie van alleen vetweefsel horen. Deze veronderstelling klopt echter niet altijd. Vooral in mammogrammen met veel klierweefsel is het onwaarschijnlijk dat de gekozen referentiewaarde alleen vetweefsel representeert. Wanneer de gekozen referentiewaarden niet bij de projectie van alleen vetweefsel hoort, wordt het klierweefselvolume en vervolgens het klierweefselpercentage onderschat. Vandaar dat we in hoofdstuk 4 naar een methode hebben gekeken om het klierweefselpercentage te bepalen. Voor mammogrammen met veel klierweefsel hebben we een andere methode ontwikkeld waarin juist het tegendeel verondersteld wordt: Het algoritme bepaald de pixelwaarde van het lichtste gedeelte in het mammogram en samen met een schatting van de hoeveelheid klierweefsel dat bij deze pixelwaarde hoort wordt een referentie berekend die zou horen bij de projectie van alleen vetweefsel. Er zijn vervolgens meerdere methodes, de ene werkt bij een lage borstdichtheid, de andere veronderstelt een hoge borstdichtheid. Een combinatie van de methodes geeft vervolgens een verbeterde schatting van het klierweefselpercentage.
Zoals eerder beschreven, neemt de sensitiviteit van de screening af naarmate de dichtheid toeneemt. Dit komt door het zogenoemde 'maskeringseffect' van borstklierweefsel. Tumor en borstklierweefsel lijken in het mammogram op elkaar waardoor het lastig is voor de radioloog om de tumor te detecteren. Om die reden wordt de kans groter dat kanker onopgemerkt blijft wanneer er veel klierweefsel aanwezig is. Vrouwen met veel klierweefsel zouden dus het meest van een screening op maat profiteren. Om te beslissen wie aanvullend onderzoek nodig heeft, moet het risico op maskering bepaald worden. In de praktijk zou het risico van alle vrouwen met een negatief mammogram (vrouwen die niet zijn doorverwezen) bepaald kunnen worden en als het risico boven een van tevoren bepaalde drempel ligt zou aanvullend onderzoek aangeboden kunnen worden. Om de voor- en nadelen in evenwicht te houden moet de risicoschatting en de drempel zodanig gekozen worden dat vooral

vrouwen voor aanvullend onderzoek in aanmerking komen bij wie de kans op aanwezigheid van een borstkanker die niet in het mammogram te zien is het grootst is. Wanneer we terugkijken naar het bevolkingsonderzoek zoals het tegenwoordig is, zijn de vrouwen met een fout negatief mammogram (de vrouwen met een intervaltumor) degene die het meest kunnen profiteren van aanvullend onderzoek. In samenhang hiermee moet wel genoemd worden, dat niet alle intervalkankers door aanvullend onderzoek detecteerbaar zijn. Er zijn ook kankers die tijdens het maken van het mammogram niet aanwezig waren en in loop van het screeningsinterval zijn ontstaan. Deze intervalkankers zijn ook met andere beeldvormende technieken niet opspoorbaar. Voor onze database weten we helaas niet welke intervalkankers er gemaskeerd waren en welke niet. Vandaar dat we veronderstellen dat alle intervalkankers gemaskeerd waren die binnen 12 maanden na de screening zijn gediagnosticeerd. De tijd om te groeien en voelbaar te worden wordt dus beperkt waardoor de kans op een tumor die tijdens het screening niet aanwezig was geminimaliseerd wordt. Vrouwen met een intervalkanker binnen 12 maanden zijn dus onze doelgroep, en we hebben gezocht naar een manier om deze vrouwen te onderscheiden van vrouwen met een terecht negatief mammogram. Want vrouwen zonder verdenking op een tumor en zonder kankerdiagnose voorafgaand aan de volgende screeningsronde zouden geen aanvullend onderzoek nodig hebben.

In hoofdstuk 5 hebben we gekeken in hoeverre BI-RADS, het percentage klierweefsel en twee andere metingen in staat zijn deze twee groepen vrouwen te onderscheiden. We waren voornamelijk geïnteresseerd in resultaten bij de transitie van BI-RADS b naar c, want dat is de classificatie die in de kliniek en ook in het bevolkingsonderzoek in de VS gebruikt wordt. Wanneer alle vrouwen met een BI-RADS categorie c of d aanvullend onderzoek aangeboden zouden krijgen, is het nodig rond 39% van de bevolking een aanvullend screeningsonderzoek aan te bieden. Wanneer vervolgens de drempel zodanig gekozen werd dat 39% van de bevolking aanvullend onderzoek krijgt, behoren significant meer vrouwen met intervalkanker tot de interventiegroep met PDA (een van de andere automatische metingen met de computer) dan met BI-RADS. Ook vergeleken met het percentage klierweefsel, zullen meer vrouwen met intervalkanker uitgenodigd zijn met PDA. Deze studie laat dus zien, dat er naast borstdichtheid ook nog andere maten zijn om het maskeringseffect te meten en dat deze misschien beter geschikt zijn voor een screening op maat.

Samenvattend zou borstdichtheid als stratificatiemethode gebruikt kunnen worden voor borstkanker screening op maat, ten eerste omdat vrouwen met een hoge borstdichtheid een verhoogd risico hebben op het ontwikkelen van borstkanker en omdat deze vrouwen ook een grotere kans lopen een intervalkanker te ontwikkelen (hoofdstuk 2). Geautomatiseerde borstdichtheidsschatting met software geeft consistentere borstdichtheidsbepalingen in opeenvolgende screeningsmammogrammen dan radiologen die de visuele BI-RADS classificatie gebruiken (hoofdstuk 3). BI-RADS wordt in de huidige praktijk het meest gebruikt voor het benaderen van de borstdichtheid, echter blijkt uit ons onderzoek dat com-

puter software beter geschikt is om het risico op een fout negatief mammogram te bepalen (hoofdstuk 5). Doordat de BI-RADS classificatie slechts uit vier klassen bestaat is de kans aanzienlijk dat het borstkankerrisico dat een vrouw loopt wordt onder- of overschat. Software geeft een volumetrische bepaling op een continue schaal die gevoeliger zou kunnen zijn voor het bepalen van de juiste screeningsmethodiek en dus een betere stratificatiemethode zou kunnen zijn. Hierbij dient men wel rekening te houden met vrouwen die op het mammogram extreem veel borstklierweefsel hebben omdat huidige volumetrische automatische methoden leiden tot een onderschatting van de borstdichtheid (hoofdstuk 4).

Desalniettemin is borstdichtheid niet de enige variabele voor screening op maat. In hoofdstuk 6 hebben we de invloed van compressie van de borst op de uitkomst van de screening onderzocht. Tijdens het maken van een mammogram word de borst gecomprimeerd, om bewegingsonscherpte, strooistraling en dosis te beperken en tevens overprojectie van borstklierweefsel te voorkomen. Elke borst word met min of meer dezelfde kracht gecomprimeerd die onafhankelijk is van borstgrootte en borstdichtheid. Vrouwen, met name degene met kleine borsten, vinden de compressie vaak erg pijnlijk. Voor sommige vrouwen is die pijnlijke ervaring de voornaamste reden om niet (meer) aan het bevolkingsonderzoek deel te nemen. Druk, gedefinieerd als kracht gedeeld door contactoppervlak met de compressieplaat, is een betere maat voor compressie omdat rekening gehouden wordt met borstgrootte, borstdichtheid en stijfheid van de borst. We hebben gekeken of er een verband bestaat tussen de sensitiviteit van het bevolkingsonderzoek en de druk die is uitgeoefend op de borst tijdens het maken van een mammogram. Verder hebben we gekeken naar het aantal (onterecht) verwijzingen, de specificiteit en het aantal tumoren. Om dit te onderzoeken hebben we eerst de vrouwen in onze dataset opgedeeld in vijf groepen op basis van de uitgevoerde druk gedurende het vervaardigen van het mammogram. Vrouwen in de eerste groep ondervonden een lage druk tijdens het maken van het mammogram, vrouwen in de vijfde groep juist een hoge druk. Uit ons onderzoek bleek dat vrouwen die in de eerste groep terecht kwamen vaker onterecht worden doorverwezen voor vervolgonderzoek, kortom een foutpositieve uitslag kregen. Deze bevinding ondersteunt dus de aanname dat een bepaalde druk en dus compressie nodig is voor een goede beeldkwaliteit en de beoordeelbaarheid van het mammogram. Vaak wordt in het bevolkingsonderzoek gedacht dat een hardere compressie altijd beter is, omdat meer compressie leidt tot betere röntgenopnames en vervolgens tot hogere detectiecijfers. Ons onderzoek laat echter zien dat bij vrouwen in de twee hoogste drukgroepen meer interval carcinomen voorkomen. Dit resulteert in lagere sensitiviteit van het bevolkingsonderzoek in de hogere drukgroepen vergeleken met de lagere drukgroepen. Dus ook een te hoge druk heeft ongewenste nadelen. Een druk gestuurde compressie, een compressie met een specifieke druk, zou zowel te veel druk en daarmee overcompressie, als ook te weinig druk en daarmee ondercompressie kunnen voorkomen. Op basis van ons onderzoek is het echter niet mogelijk een streefdruk te bepalen.

Zusammenfassung

Sowohl in den Niederlanden als auch in Deutschland, sowie in vielen anderen westlichen Ländern, werden symptomfreie Frauen eingeladen alle zwei Jahre an dem Brustkrebsfrüherkennungsprogramm teilzunehmen. Man bietet das Screening an, um den Brustkrebs möglichst früh zu erkennen und die Heilungschancen, im Falle einer Erkrankung, zu erhöhen, denn Brustkrebs ist die häufigste Krebsdiagnose bei Frauen. Im Screening werden von jeder Brust zwei Röntgenaufnahmen erstellt. Diese werden dann von Radiologen im Bezug auf eine mögliche Abnormalität beurteilt. Auch wenn mehrere Studien die Effektivität von Brustkrebsfrüherkennungsprogrammen bewiesen haben, hat das Screening nicht nur Vorteile. Zum einen werden mehr Frauen als nötig über einen möglichen Befund informiert, welcher sich nach weiteren Untersuchungen als gutartig oder als normales Gewebe herausstellt (falsch positiver Befund). Zum anderen werden nicht alle Tumore innerhalb des Früherkennungsprogrammes entdeckt. Ungefähr 16-32% der Tumore werden außerhalb des Screenings diagnostiziert.

Das Früherkennungsprogramm in seiner heutigen Form ist nicht für alle Frauen gleichermaßen effektiv. Daher wird über stratifiziertes Screening nachgedacht. Das Screeningprogramm würde dann von dem Brustkrebsrisiko und dem Risiko, dass der Tumor nicht mit Mammographie innerhalb des Screenings entdeckt werden könnte, abhängen. Anstelle eines Mammogrammes alle zwei Jahre, können anderen bildgebende Verfahren, zum Beispiel Ultraschall oder MRT, eingesetzt werden. Auch könnte das Screeningintervall auf die individuellen Bedürfnisse angepasst werden.

Noch ist nicht klar welche Parameter zur Stratifizierung verwendet werden können. Eine Stratifizierung auf Basis der Brustdichte ist möglich. Häufig wird die Brustdichte, die Menge Drüsengewebe im Verhältnis zur Brustgröße, von Radiologen geschätzt. Diese verwenden dazu den BI-RADS Atlas. In diesem wird die Brustdichte in vier Kategorien unterteilt. Problem der BI-RADS Kategorien ist jedoch die Inter- und Intra-Leser Variabilität. Die zugeordnete Kategorie variiert sowohl zwischen Radiologen als auch bei der Beurteilung der gleichen Person zu verschiedenen Zeitpunkten. Zudem stehen mehrere Algorithmen zur Brustdichtemessung zur Verfügung. Die meisten Algorithmen unterscheiden dazu Drüsengewebe von Fettgewebe. Sowohl die absolute Menge Drüsengewebe (in cm^3) als auch die Menge Drüsengewebe im Verhältnis zur Brustgröße können zur Stratifizierung herrangezogen werden. Die automatisierte Messung der Brustdichte wurde durch die Einführung der digitalen Mammographie vereinfacht.

Dass die Mammographie nicht für alle Frauen gleichermaßen geeignet ist, wird in Kapitel 2 sichtbar. Neben verschiedenen anderen Parametern wurde die Sensitivität, der Anteil im Screening diagnostizierte Tumore im Vergleich zu allen Brustkrebserkrankungen, in Bezug auf die Brustdichte gemessen. Wir fanden heraus, dass die Sensitivität abnimmt, wenn die Brustdichte zunimmt. Zudem ist sichtbar, dass das Risiko an Brustkrebs zu erkranken mit Zunahme der Brustdichte ansteigt. Frauen mit hoher Brustdichte haben somit nicht nur ein erhöhtes Brustkrebsrisiko, sondern auch ein erhöhtes Risiko, dass der Tumor nicht inner-

halb des Früherkennungsprogramms erkannt wird. Diese Frauen würden somit am stärksten von einem stratifizierten Früherkennungsprogramm profitieren.
In den letzten Jahren wurde in den USA auch außerhalb der Wissenschaft über Brustdichte diskutiert. Die “breast density laws” verpflichten Radiologen nun Frauen über ihre Brustdichte und die damit einhergehenden Risiken zu informieren. Häufig unterscheidet man lediglich Frauen mit viel Drüsengewebe (BI-RADS c und d) von solchen mit wenig Drüsengewebe (BI-RADS a und b). Wenn man eine solche Klassifizierung zur Stratifizierung in verschiedene Screeningregime verwenden will, ist es wichtig, dass die Klassifizierung konsistent ist. Nicht nur im Bezug auf Inter- und Intra-Leser Reliabilität, sondern auch im Bezug auf temporelle Daten. Die Klassifizierung sollte sich nicht alle zwei Jahre, mit jeder neuen Aufnahme, ändern.
In Kapitel 3 wird die Klassifizierung in vier, beziehungsweise zwei, Gruppen von temporellen Daten untersucht. Dazu wurden 500 Mammogrammpaare sowohl von Radiologen mit Hilfe des BI-RADS Atlas beurteilt als auch von vollautomatischer Software auf ihre Brustdichte hin untersucht. Die Software (Volpara) bestimmt dazu das Volumen des Drüsengewebes und das Brustvolumen. Anschließend wird der Prozentsatz Drüsengewebe bestimmt, welcher zur Unterteilung in vier Kategorien (VDG) genutzt wird. Wir fanden heraus, dass in 86-91% der Fälle beide Aufnahmen eines Paares zur gleichen Gruppe gehörten. Zudem wurde untersucht wie konsistent die Dichtemessung ist, wenn die Mammogramme eines Paares von verschieden Radiologen beurteilt wurden, wie es im Screening der Fall ist. Es zeigte sich, dass die Messung mit der automatischen Software konsistenter als die Simulation der Screeningsituation ist. Dies legt daher den Schluß nahe, dass die Software besser geeigent ist die Brustdichte zu bestimmen.
Der BI-RADS Atlas ermöglicht eine Klassifizierung der Brustdichte in vier Kategorien. Mit Software ist es jedoch möglich, eine kontinuierliche Bestimmung der Brustdichte zu erhalten. Die volumetrische Brustdichte Messung auf Basis eines Mammogrammes setzt verschiedene Annahmen vorraus. Schließlich wird ein Volumen auf Basis eines zweidimensionalen Bildes berechnet. Einer der bestehenden Algorithmen verwendet für jede Aufnahme einen internen Referenzwert um die Brustdichte zu bestimmen. Dieser Referenzwert ist repräsentativ für die dunkelste Stelle innerhalb der segmentierten Brust. Es wird angenommen, dass dieser Pixelwert zu der Projektion von Fettgewebe gehört und dass sich kein Drüsengewebe zwischen Strahlenquelle und dem Detektor befand. Diese Annahme stimmt jedoch nicht, wenn man eine Aufnahme von einer Brust mit viel Drüsengewebe betrachtet. Studien haben gezeigt, dass in diesem Fall die Brustdichte unterschätzt wird. In diesen Studien wird die Brustdichtemessung in Mammographie-Aufnahmen mit der Brustdichtemessung in MRT Aufnahmen verglichen. Die dreidimensionale MRT Aufnahme ermöglicht eine genaue Bestimmung der Brustdichte in dem jedes Voxel (3D Version eines Pixels) als Drüsengewebe oder Fettgewebe klassifiziert wird. Wenn MRT und Mammogramm innerhalb weniger Wochen voneinander aufgenommen wurden, nimmt man an, dass die jeweils

gemessene Brustdichte gleich sein sollte.
In Kapitel 4 werden drei Methoden zur Referenzwertbestimmung präsentiert. In den ersten beiden Methoden wird der Pixelwert, welcher zu der dunkelsten Stelle in der Aufnahme gehört, und somit representativ für die Projektion von aussschließlich Fettgewebe sein soll, bestimmt. Dabei scheint die zweite Methode besser geeignet zu sein für Aufnahmen mit mehr Drüsengewebe während die erste Methode gute Resultate liefert für Aufnahmen mit sehr wenig Drüsengewebe. Die dritte Methode verwendet einen Pixelwert, der zu einer der hellsten Regionen im Mammogramm gehört, um den Referenzwert zu bestimmen der zu der Projektion von Fettgewebe gehören würde und nicht in einem Mammogramm mit sehr viel Drüsengewebe vorhanden ist. Diese alternative Methode ist nötig, wenn der Referenzwert bestimmt mit den ersten Methoden nicht representativ für die Projektion von Fettgewebe ist. Da die erste Methode sehr gut bei Mammogrammen mit wenig Drüsengewebe funktioniert und die anderen beiden Methoden entwickelt wurden unter der Annahme, dass viel Drüsengewebe vorhanden ist, liefert eine Kombination der Methoden die besten Resultate, im Vergleich zu der Dichtemessung der MRT Aufnahmen.
In Kapitel 2 zeigten wir, dass die Sensitivität mit Zunahme der Brustdichte abnimmt und dass das Brustkrebsrisiko zunimmt. Auf Grund dieser Gegebenheiten bekommen Frauen in manchen Staaten in den USA schon zusätzliche bildgebenden Verfahren angeboten. Derzeit entscheidet lediglich die Brustdichte, geschätzt mit BI-RADS oder Software, wer für zusätzliches Screening in Frage kommt. Das Risiko auf "Maskierung" wird noch nicht evaluiert. Der "Maskierungseffekt" beschreibt die Tatsache, dass Tumorgewebe und Drüsengewebe auf dem Mammogramm gleich aussehen, und dass im Falle von viel Drüsengewebe es nicht möglich ist die beiden voneinander zu unterscheiden. Der Tumor hat sozusagen eine "Maske" auf und verschleiert sein wahres Gesicht. Durch die Maskierung wird der Tumor nicht innerhalb des Screenings sondern stattdessen zwischen zwei Screeningrunden entdeckt. Auch wenn das Maskierungsrisiko mit Zunahme der Brustdichte zunimmt, heißt dies nicht, dass Brustdichte das beste Maß für den Maskierungseffekt ist. Um den Maskierungseffekt zu messen, haben wir in Kapitel 5 zurecht negative Mammogrammen mit falsch negativen Mammogrammen (Mammogrammen von Frauen, welche nicht durchverwiesen wurden und welche dennoch innerhalb von 12 Monaten nach den Aufnahmen eine Brustkrebsdiagnose erhielten) verglichen. Das ideale Maß für den Maskierungseffekt würde alle Frauen, die später einen Intervalltumor diagnostiziert bekommen, identifizieren. Durch die Annahme, dass der Tumor schon vorhanden ist, würden diese Frauen am meisten von zusätzlichen Screeningmaßnahmen profitieren. In besagtem Kapitel haben wir verschiedene Maße auf die Eigenschaft hin verglichen, falsch negative Mammogramme von zurecht negativen Aufnahmen zu trennen. Zum einen wurden alle Mammogramme von einem Radiologen mit Hilfe des BI-RADS Atlas beurteilt, auch wurde in allen Mammogrammen die Brustdichte mit Software gemessen. Zwei weitere präsentierte Methoden verwenden eine Dichtekarte. Die Dichtekarte ist vergleichbar mit einer Höhenkarte und gibt für jede Positi-

on in dem Mammogramm die Menge von Drüsengewebe (in cm) an. Das erste Maß ist PDA, der Prozentsatz der Pixel mit einem Drüsengewebeanteil von mehr als 1 cm. Das zweite Maß (DTMM) verwendete die Dichtekarte in Kombination mit der Tumorgrößenverteilung innerhalb des Screenings. Letztere Methode verwendet unter anderem das Wissen, dass das Risiko zunimmt, wenn die Menge Drüsengewebe zunimmt und dass das Risiko auf Maskierung abnimmt, wenn die Tumorgröße zunimmt. Vorallem die automatisierte Messung PDA ist im Vergleich zur Klassifizierung mit BI-RADS besser geeignet um Frauen mit zukünftigem Intervalltumor von Frauen mit zurecht negativem Befund zu unterscheiden, angenommen, dass Frauen mit der BI-RADS Kategorien c und d ein erhöhtes Risiko haben und dass dementsprechend ungefähr 38.5% der Frauen für zusätzliche bildgebende Verfahren in Betracht kommen.

Um die Strahlenbelastung zu verringern, Gewebeschichten zu trennen und Bewegungseffekte zu vermeiden, wird die Brust während der Mammographie zusammengedrückt. Die Kompression wird von vielen Frauen als schmerzhaft empfunden und ist einer der häufigsten Gründe nicht am Screening teilzunehmen. Ein Maß für die Kompression ist Kraft, welche auch in den Metainformationen einer jeden Aufnahme gespeichert wird. Die Vorgehensweise bei der Brustkompression ist hinsichtlich der Kompressionskraft jedoch nicht standardisiert. Ein besseres Maß für die Kompression könnte Druck sein. Druck ist definiert als Kraft geteilt durch Oberfläche. Um den Druck der Brustkompression zu berechnen dividiert man die gemessene Kraft durch die Kontaktoberfläche, die Fläche in der die Brust die Kompressionsplatte berührt. Diese kann mit Hilfe von Software berechnet werden. Während die Kraft unabhängig von den Gegebenheiten der Brust ist, wird in der Druckmessung die Brustgröße und Dichte einbezogen. Eine große Brust, die vorallem aus Fettgewebe besteht, nimmt eine gewisse Kraft anders wahr als eine kleine Brust. In Kapitel 6 wird die Qualität des Screeningprogrammes im Hinblick auf Kompression beurteilt. Dazu wurden die Mammogramme auf fünf Gruppen verteilt, sodass jede Gruppe gleich viele Mammogramme beinhaltete und der Druck mit jeder Gruppe zunahm. Anschließend wurde für jede Gruppe die Sensitivität, Verweißrate, Inzidenzraten und ähnliche Parameter berechnet. Es scheint, dass in der ersten Gruppe, welche Mammograme mit geringem Druck beinhaltet, Frauen häufiger über einen möglichen Befund informiert werden, der sich anschließend als falscher Alarm herausstellt. Dies zeigt, dass eine gewisse Kompression notwendig ist um verschiedene Gewebeschichten ausreichend zu trennen und eine gute Bildqualität zu gewährleisten. Andererseits nimmt die Sensitivität in den letzten beiden Gruppen ab. Eine zu starke Kompression scheint den Kontrast zwischen verschiedenen Geweben zu verringern wodurch Tumore weniger verdächtig erscheinen. Dies hängt vielleicht mit einer reduzierten Blutzufuhr zusammen. Vielleicht gibt es einen Optimaldruck, der sowohl eine zu niedrige als auch zu starke Kompression verhindert. Weitere Forschungen auf diesem Gebiet sind dazu notwendig.

Zusammenfassend: Brustdichte ist ein bekannter Brustkrebsrisikofaktor. Frauen mit hoher Brustdichte haben ein erhöhtes Brustkrebsrisiko und ein erhöhtes Risiko, dass der Tumor nicht innerhalb des Brustkrebsfrüherkennungsprogrammes erkannt wird. Diese Frauen würden daher am stärksten von einem stratifizierten Screening profitieren. Die Brustdichte kann auf verschiedene Weisen gemessen werden. Dabei gilt es zu beachten, dass der Prozentsatz Drüsengewebe leicht unterschätzt werden kann, vorallem in Aufnahmen mit viel Drüsengewebe. Anstelle einer kontinuierlichen Messung der Brustdichte können auch Brustdichtegruppen verwendet werden. Eine Klassifizierung mit Software verhindert Inter- und Intra-Leser Variabilität und klassifiziert temporelle Daten konsistenter im Vergleich zu der Beurteilung des Radiologen. Auch wenn die Brustdichtemessung inzwischen weit verbreitet ist, andere Messungen sind möglicherweise besser geeignet Mammogramme von Frauen mit zukünftigen Intervalltumoren von zurecht negativen Aufnahmen zu unterscheiden, als die hier untersuchten Brustdichtegruppen und der Prozentsatz Drüsengewebe. Die Einführung eines Idealdrucks könnte zudem vielleicht unnötige Überweisungen und Intervalltumore verhindern.

Publications

Papers in international journals

J.O.P. Wanders, **K. Holland**, N. Karssemeijer, P.H.M. Peeters, W.B. Veldhuis, R.M. Mann and C.H. van Gils. "Changes in volumetric breast density and the association with breast cancer risk", to be submitted.

J.O.P Wanders, C.H. van Gils, N. Karssemeijer, **K. Holland**, M. Kallenberg, P.H.M. Peeters, M. Nielsen and M. Lillholm. "The combined effect of mammographic texture and density on breast cancer risk", submitted.

K. Holland, I. Sechopoulos, R.M. Mann, G.J. den Heeten, C.H. van Gils and N. Karssemeijer. "Influence of breast compression pressure on the performance of population-based mammography screening", submitted.

J.O.P Wanders, **K. Holland**, N. Karssemeijer, P.H.M. Peeters, W.B. Veldhuis, R.M. Mann and C.H. van Gils. "The effect of volumetric breast density on the risk of screen detected and interval breast cancers: a cohort study", Breast Cancer Research, 2017.

K. Holland, A. Gubern-Mérida, R.M. Mann and N. Karssemeijer. "Optimization of volumetric breast density estimation in digital mammograms", Physics in Medicine and Biology, 2017.

K. Holland, C.H. van Gils, R.M. Mann and N. Karssemeijer. "Quantification of masking risk in screening mammography with volumetric breast density maps", Breast Cancer Research and Treatment, 2017.

J.O.P. Wanders, **K. Holland**, W.B. Veldhuis, R.M. Mann, R.M. Pijnappel, P.H.M. Peeters, C.H. van Gils and N. Karssemeijer. "Volumetric breast density affects performance of digital screening mammography", Breast Cancer Research and Treatment, 2017.

M.U. Dalmis, G. Litjens, **K. Holland**, R. Mann, N. Karssemeijer and A. Gubern-Mérida. "Using Deep Learning to Segment Breast and Fibroglandular Tissue in MRI Volumes", Medical Physics, 2017.

K. Holland, J. van Zelst, G.J. den Heeten, M. Imhof-Tas, R.M. Mann, C.H. van Gils and N. Karssemeijer. "Consistency of breast density categories in serial screening mammograms: A comparison between automated and human assessment", The Breast, 2016.

M. Kallenberg, K. Petersen, M. Nielsen, A. Ng, P. Diao, C. Igel, C. Vachon, **K. Holland**, N. Karssemeijer and M. Lillholm. "Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring", IEEE Transactions on Medical Imaging, 2016.

Papers in conference proceedings

K. Holland, I. Sechopoulos, G. den Heeten, R.M. Mann and N. Karssemeijer. "Performance of breast cancer screening depends on mammographic compression", Breast Imaging - 13th International Workshop of Breast Imaging (IWDM), 2016.

M. Kallenberg, M. Nielsen, **K. Holland**, N. Karssemeijer, C. Igel and M. Lillholm. "Learning Density Independent Texture Features", Breast Imaging - 13th International Workshop of Breast Imaging (IWDM), 2016.

K. Holland, C.H. van Gils, J.O.P. Wanders, R.M. Mann and N. Karssemeijer. "Quantification of mammographic masking risk with volumetric breast density maps: How to select women for supplemental screening", Medical Imaging, Proceedings of the SPIE, 2016.

K. Holland, M. Kallenberg, R. Mann, C. van Gils and N. Karssemeijer. "Stability of Volumetric Tissue Composition Measured in Serial Screening Mammograms", Breast Imaging - 12th International Workshop of Breast Imaging (IWDM), 2014.

Abstracts in conference proceedings

N. Karssemeijer, **K. Holland**, I. Sechopoulos, R.M. Mann, G.J. den Heeten and C.H. van Gils. "High breast compression in mammography may reduce sensitivity", Annual Meeting of the Radiological Society of North America, 2016.

M. Kallenberg, M. Nielsen, **K. Holland**, N. Karssemeijer and M. Lillholm. "Breast cancer risk prediction with density independent texture features", Annual Meeting of the Radiological Society of North America, 2016.

J.O.P. Wanders, **K. Holland**, P.H.M. Peeters, N. Karssemeijer and C.H. van Gils. "Volumetric breast density and the risk of screen detected and interval breast cancer", WEON - The annual conference of the Dutch Epidemiological Society, 2016.

J.O.P Wanders, **K. Holland**, P.H.M. Peeters, N. Karssemeijer and C.H. van Gils. "Volumetric breast density and the risk of screen detected and interval breast cancer", Annual conference of the International Agency for Research on Cancer, 2016.

K. Holland, C.H. van Gils, J.O.P. Wanders, R.M. Mann and N. Karssemeijer. "How can we identify women at risk for a masked cancer, who may benefit from supplemental screening?", Annual Meeting of the Radiological Society of North America, 2015.

K. Holland, C.H. van Gils, J.O.P. Wanders, R.M. Mann and N. Karssemeijer. "Optimisation of the selection of women with an increased risk of a masked tumour for supplementary screening", Annual Meeting of the Radiological Society of North America, 2015.

K. Holland, C.H. van Gils, J.O.P. Wanders, R.M. Mann and N. Karssemeijer. "Consistency of density categories over multiple screening rounds using volumetric breast density", Annual Meeting of the Radiological Society of North America, 2015.

M. Kallenberg, M. Lillholm, P. Diao, K. Petersen, **K. Holland**, N. Karssemeijer, C. Igel and M. Nielsen. "Assessing Breast Cancer Masking Risk with Automated Texture Analysis in Full Field Digital Mammography", Annual Meeting of the Radiological Society of North America, 2015.

K. Holland, A. Gubern-Mérida, R.M. Mann and N. Karssemeijer. "Improved volumetric breast density assessment in dense breasts", 7th International Workshop on Breast Densitometry and Cancer Risk Assessment, 2015.

M. Kallenberg, M. Lillholm, P. Diao, **K. Holland**, N. Karssemeijer, C. Igel and M. Nielsen. "Assessing breast cancer masking risk in full field digital mammography with automated texture analysis", 7th International Workshop on Breast Densitometry and Cancer Risk Assessment, 2015.

J.O.P. Wanders, **K. Holland**, P.H.M. Peeters, N. Karssemeijer and C.H. van Gils. "Combined effect of dense and nondense breast volume on breast cancer risk", 7th International Workshop on Breast Densitometry and Cancer Risk Assessment, 2015.

J.O.P. Wanders, **K. Holland**, P.H.M. Peeters, N. Karssemeijer and C.H. van Gils. "Volumetric breast density and the risk of screen detected and interval breast cancer", 7th International Workshop on Breast Densitometry and Cancer Risk Assessment, 2015.

J.O.P. Wanders, **K. Holland**, W.B. Veldhuis, R.M. Mann, P.H.M. Peeters, C.H. van Gils and N. Karssemeijer. ”Effect of volumetric mammographic density on performance of a breast cancer screening program using full-field digital mammography”, 7th International Workshop on Breast Densitometry and Cancer Risk Assessment, 2015.

J.O.P. Wanders, **K. Holland**, W.B. Veldhuis, R.M. Mann, P.H.M. Peeters, C.H. van Gils and N. Karssemeijer. ”Effect of mammographic density on performance of a breast cancer screening program using full-field digital mammography”, European Congress of Epidemiology, 2015

M.G. Kallenberg, K. Petersen, M. Lillholm, D.R. Jørgensen, P. Diao, **K. Holland**, N. Karssemeijer, C. Igel and M. Nielsen. ”Automated texture scoring for assessing breast cancer masking risk in full field digital mammography”, European Congress of Radiology, 2015.

J.O.P. Wanders, **K. Holland**, W. Veldhuis, R. Mann, P.H.M. Peeters, C.H. van Gils and N. Karssemeijer. ”Effect of volumetric mammographic density on performance of a breast cancer screening program using full-field digital mammography”, European Congress of Radiology, 2015.

J.O.P. Wanders, N. Karssemeijer, **K. Holland**, P.H.M. Peeters and C.H. van Gils. ”Breast density: the size of the problem in Dutch screening participants”, WEON - The annual conference of the Dutch Epidemiological Society, 2014

M.G.J. Kallenberg, **K. Holland**, J.O.P. Wanders, C.H. van Gils, and N. Karssemeijer. ”Association between Automated, Volumetric Breast Density Measures and Breast Cancer in a Large Screening Population”, 6th International Workshop on Breast Densitometry and Cancer Risk Assessment, 2013.

Bibliography

[1] J. Ferlay, E. Steliarova-Foucher, J. Lortet-Tieulent, S. Rosso, J W W. Coebergh, H. Comber, D. Forman, and F. Bray, 'Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012'. European Journal of Cancer, 2013, 49, 1374–1403, doi:10.1016/j.ejca.2012.12.027.

[2] M. Mistry, D.M. Parkin, A.S. Ahmad, and P. Sasieni, 'Cancer incidence in the United Kingdom: projections to the year 2030'. British Journal of Cancer, 2011, 105 (11), 1795–1803, doi:10.1038/bjc.2011.430.

[3] H.K. Weir, T.D. Thompson, A. Soman, B. Møller, and S. Leadbetter, 'The past, present, and future of cancer incidence in the United States: 1975 through 2020'. Cancer, 2015, 121 (11), 1827–1837, doi:10.1002/cncr.29258.

[4] A. Antoniou, P.D.P. Pharoah, S. Narod, H.A. Risch, J.E. Eyfjord, J.L. Hopper, N. Loman, H. Olsson, O. Johannsson, A. Borg, B. Pasini, P. Radice, S. Manoukian, D.M. Eccles, N. Tang, E. Olah, H. Anton-Culver, E. Warner, J. Lubinski, J. Gronwald, B. Gorski, H. Tulinius, S. Thorlacius, H. Eerola, H. Nevanlinna, K. Syrjäkoski, O-P. Kallioniemi, D. Thompson, C. Evans, J. Peto, F. Lalloo, D.G. Evans, and D.F. Easton, 'Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: A combined analysis of 22 studies'. The American Journal of Human Genetics, 2003, 72 (5), 1117–1130, doi:10.1086/375033.

[5] P.D. Pharoah, N.E. Day, S. Duffy, D.F. Easton, and B.A. Ponder, 'Family history and the risk of breast cancer: a systematic review and meta-analysis'. International Journal of Cancer, 1997, 71 (5), 800–809, doi:10.1002/(SICI)1097-0215(19970529)71:5<800::AID-IJC18>3.3.CO;2-R.

[6] M. Ewertz, S.W. Duffy, H.O. Adami, G. Kvåle, E. Lund, O. Meirik, A. Mellemgaard, I. Soini, and H. Tulinius, 'Age at first birth, parity and risk of breast cancer: A meta-analysis of 8 studies from the nordic countries'. International Journal of Cancer, 1990, 46 (4), 597–603, doi:10.1002/ijc.2910460408.

[7] Collaborative Group on Hormonal Factors in Breast Cancer, 'Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease'. The Lancet, 2002, 360 (9328), 187–195, doi:10.1016/S0140-6736(02)09454-0.

[8] Collaborative Group on Hormonal Factors in Breast Cancer, 'Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies'. The Lancet Oncology, 2012, 13 (11), 1141–1151, doi:10.1016/S1470-2045(12)70425-4.

[9] Collaborative Group on Hormonal Factors in Breast Cancer, 'Alcohol, tobacco and breast cancer–collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease'. British Journal of Cancer, 2002, 87 (11), 1234–1245, doi:10.1038/sj.bjc.6600596.

[10] V.A. McCormack and I. dos Santos Silva, 'Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis'. Cancer Epidemiology, Biomarkers and Prevention, 2006, 15, 1159–1169, doi:10.1158/1055-9965.EPI-06-0034.

[11] World Health Organization Classification of Tumours, F.A. Tavassoli, and L.M. Roth, Pathology and genetics of tumours of the breast and female genital organs. 2003.

[12] S. Hofvind, A. Ponti, J. Patnick, N. Ascunce, S. Njor, M. Broeders, L. Giordano, A. Frigerio, and S. Törnberg, 'False-positive results in mammographic screening for breast cancer in Europe: a literature review and survey of service screening programmes'. Journal of Medical Screening, 2012, 19, 57–66, doi:10.1258/jms.2012.012083.

[13] Rijksinstituut voor Volksgezondheid en Milieu, 'Onderzoek naar borstkanker, Information brochures for women'. 2016.

[14] S. Törnberg, L. Kemetli, N. Ascunce, S. Hofvind, A. Anttila, B. Sèradour, E. Paci, C. Guldenfels, E. Azavedo, A. Frigerio, V. Rodrigues, and A. Ponti, 'A pooled analysis of interval cancer rates in six European countries'. European Journal of Cancer Prevention, 2010, 19 (2), 87–93, doi:10.1097/CEJ.0b013e32833548ed.

[15] J. Nederend, L.E. Duijm, A.C. Voogd, J.H. Groenewoud, F.H. Jansen, and M.W Louwman, 'Trends in incidence and detection of advanced breast cancer at biennial screening mammography in the Netherlands: a population based study'. Breast Cancer Research, 2012, 14 (1), R10, doi:10.1186/bcr3091.

[16] L. Domingo, D. Salas, R. Zubizarreta, M. Baré, G. Sarriugarte, T. Barata, J. Ibáñez, J. Blanch, M. Puig-Vives, A. Fernández, X. Castells, M. Sala, and I. N. C. A Study Group , 'Tumor phenotype and breast density in distinct categories of interval cancer: results of population-based mammography screening in Spain'. Breast Cancer Research, 2014, 16 (1), R3, doi:10.1186/bcr3595.

[17] J. Holm, K. Humphreys, J. Li, A. Ploner, A. Cheddad, M. Eriksson, S. Törnberg, P. Hall, and K. Czene, 'Risk factors and tumor characteristics of interval cancers by mammographic density'. Journal of Clinical Oncology, 2015, 33 (9), 1030–1037, doi:10.1200/JCO.2014.58.9986.

[18] B. Meshkat, R.S. Prichard, Z. Al-Hilli, G.A. Bass, C. Quinn, A. O'Doherty, J. Rothwell, J. Geraghty, D. Evoy, and E.W. McDermott, 'A comparison of clinicopathological characteristics between symptomatic and interval breast cancer'. The Breast, 2015, 24 (3), 278–282, doi:10.1016/j.breast.2015.02.032.

[19] B.C. Yankaskas, M.J. Schell, R.E. Bird, and D.A. Desrochers, 'Reassessment of breast cancers missed during routine screening mammography'. American Journal of Roentgenology, 2001, 177 (3), 535–541, doi:10.2214/AJR.177.3.1770535.

[20] R.F. Brem, J. Baum, M. Lechner, S. Kaplan, S. Souders, L.G. Naul, and J. Hoffmeister, 'Improvement in sensitivity of screening mammography with computer-aided detection: A multiinstitutional trial'. American Journal of Roentgenology, 2003, 181 (3), 687–693, doi:10.2214/AJR.181.3.1810687.

[21] N. Karssemeijer, J.D.M. Otten, A.L.M. Verbeek, J.H. Groenewoud, H.J. de Koning, J.H.C.L. Hendriks, and R. Holland, 'Computer-aided detection versus independent double reading of masses on mammograms'. Radiology, 2003, 227 (1), 192–200, doi:10.1148/radiol.2271011962.

[22] C.M. Vachon, C.H. van Gils, T.A. Sellers, K. Ghosh, S. Pruthi, K.R. Brandt, and V.S. Pankratz, 'Mammographic density, breast cancer risk and risk prediction'. Breast Cancer Research, 2007, 9, 217, doi:10.1186/bcr1829.

[23] N.F. Boyd, L.J. Martin, M.J. Yaffe, and S. Minkin, 'Mammographic density and breast cancer risk: current understanding and future prospects'. Breast Cancer Research, 2011, 13 (6), 223, doi:10.1186/bcr2942.

[24] A. Eng, Z. Gallant, J. Shepherd, V.A. McCormack, J. Li, M. Dowsett, S. Vinnicombe, S. Allen, and I. Dos-Santos-Silva, 'Digital mammographic density and breast cancer risk: a case-control study of six alternative density assessment methods'. Breast Cancer Research, 2014, 16 (5), 439, doi:10.1186/s13058-014-0439-1.

[25] C.M. Checka, J.E. Chun, F.R. Schnabel, J. Lee, and H. Toth, 'The relationship of mammographic density and age: Implications for breast cancer screening'. American Journal of Roentgenology, 2012, 198 (3), W292–W295, doi:10.2214/AJR.10.6049.

[26] N.F. Boyd, H. Guo, L.J. Martin, L. Sun, J. Stone, E. Fishell, R.A. Jong, G. Hislop, A. Chiarelli, S. Minkin, and M.J. Yaffe, 'Mammographic density and the risk and detection of breast cancer'. New England Journal of Medicine, 2007, 356 (3), 227–236, doi:10.1056/NEJMoa062790.

[27] K. Kerlikowske, 'The mammogram that cried Wolfe'. New England Journal of Medicine, 2007, 356, 297–300, doi:10.1056/NEJMe068244.

[28] E.D. Pisano, R.E. Hendrick, M.J. Yaffe, J.K. Baum, S. Acharyya, J.B. Cormack, L.A. Hanna, E.F. Conant, L.L. Fajardo, L.W. Bassett, C.J. D'Orsi, R.A. Jong, M. Rebner, A.N.A. Tosteson, C.A. Gatsonis, and D. M. I. S. T Investigators Group , 'Diagnostic accuracy of digital versus film mammography: Exploratory analysis of selected population subgroups in DMIST'. Radiology, 2008, 246 (2), 376–383, doi:10.1148/radiol.2461070200.

[29] K. Kerlikowske, W. Zhu, A.N.A. Tosteson, B.L. Sprague, J.A. Tice, C.D. Lehman, D.L. Miglioretti, and Breast Cancer Surveillance Consortium , 'Identifying women with dense breasts at high risk for interval cancer: a cohort study'. Annals of Internal Medicine, 2015, 162 (10), 673–681, doi:10.7326/M14-1465.

[30] S. Destounis, L. Johnston, R. Highnam, A. Arieno, R. Morgan, and A. Chan, 'Using volumetric breast density to quantify the potential masking risk of mammographic density'. American Journal of Roentgenology, 2017, 208 (1), 222–227, doi:10.2214/AJR.16.16489.

[31] M.V. Prummel, D. Muradali, R. Shumak, V. Majpruz, P. Brown, H. Jiang, S.J. Done, M.J. Yaffe, and A.M. Chiarelli, 'Digital compared with screen-film mammography: Measures of diagnostic accuracy among women screened in the Ontario Breast Screening Program'. Radiology, 2016, 278 (2), 365–373, doi:10.1148/radiol.2015150733.

[32] S. Weigel, W. Heindel, J. Heidrich, H.-W. Hense, and O. Heidinger, 'Digital mammography screening: sensitivity of the programme dependent on breast density.' European radiology, 2017, 27, 2744–2751, doi:10.1007/s00330-016-4636-4.

[33] J.T. Schousboe, K. Kerlikowske, A. Loh, and S.R. Cummings, 'Personalizing mammography by breast density and other risk factors for breast cancer: analysis of health benefits and cost-effectiveness'. Annals of Internal Medicine, 2011, 155, 10–20, doi:10.1059/0003-4819-155-1-201107050-00003.

[34] J.M. Ho, N. Jafferjee, G.M. Covarrubias, M. Ghesani, and B. Handler, 'Dense breasts: a review of reporting legislation and available supplemental screening options'. American Journal of Roentgenology, 2014, 203 (2), 449–456, doi:10.2214/AJR.13.11969.

[35] J.N. Wolfe, 'Breast patterns as an index of risk for developing breast cancer'. American Journal of Roentgenology, 1976, 126 (6), 1130–1137, doi:10.2214/ajr.126.6.1130.

[36] N. F. Boyd, B. O'Sullivan, J. E. Campbell, E. Fishell, I. Simor, G. Cooke, and T. Germanson, 'Mammographic signs as risk factors for breast cancer'. British Journal of Cancer, 1982, 45 (2), 185–193, doi:10.1038/bjc.1982.32.

[37] C.J. D'Orsi, L.W. Bassett, W.A. Berg, S.A. Feig, V.P. Jackson, and D.B. Kopans, Breast Imaging Reporting and Data System (BI-RADS) Atlas. Reston, VA, 4 edition, 2003.

[38] C.J. D'Orsi, E.A. Sickles, E.B. Mendelson, and E.A. Morris et al., ACR BI-RADS Atlas, Breast Imaging Reporting and Data System. Reston, VA, 5 edition, 2013.

[39] K. Kerlikowske, D. Grady, J. Barclay, S.D. Frankel, S.H. Ominsky, E.A. Sickles, and V. Ernster, 'Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System'. JNCI Journal of the National Cancer Institute, 1998, 90 (23), 1801–1809, doi:10.1093/jnci/90.23.1801.

[40] S. Ciatto, N. Houssami, A. Apruzzese, E. Bassetti, B. Brancato, F. Carozzi, S. Catarzi, M. P. Lamberini, G. Marcelli, R. Pellizzoni, B. Pesce, G. Risso, F. Russo, and A. Scorsolini, 'Categorizing breast mammographic density: intra- and interobserver reproducibility of BI-RADS density categories'. The Breast, 2005, 14 (4), 269–275, doi:10.1016/j.breast.2004.12.004.

[41] E.A. Ooms, H.M. Zonderland, M.J.C. Eijkemans, M. Kriege, B. Mahdavian Delavary, C.W. Burger, and A.C. Ansink, 'Mammography: Interobserver variability in breast density assessment'. The Breast, 2007, 16, 568–576, doi:10.1016/j.breast.2007.04.007.

[42] C.C. Gard, E.J. Aiello Bowles, D.L. Miglioretti, S.H. Taplin, and C.M. Rutter, 'Misclassification of Breast Imaging Reporting and Data System (BI-RADS) mammographic density and Implications for breast density reporting legislation'. The Breast Journal, 2015, 21 (5), 481–489, doi:10.1111/tbj.12443.

[43] E.U. Ekpo, U.P. Ujong, C. Mello-Thoms, and M.F. McEntee, 'Assessment of interradiologist agreement regarding mammographic breast density classification using the fifth edition of the BI-RADS atlas'. American Journal of Roentgenology, 2016, 206 (5), 1119–1123, doi:10.2214/AJR.15.15049.

[44] A. Irshad, R. Leddy, S. Ackerman, A. Cluver, D. Pavic, A. Abid, and M.C. Lewis, 'Effects of changes in BI-RADS density assessment guidelines (fourth versus fifth edition) on breast density assessment: Intra- and interreader agreements and density distribution'. American Journal of Roentgenology, 2016, 207 (6), 1366–1371, doi:10.2214/AJR.16.16561.

[45] R. Highnam and M. Brady, Mammographic Image Analysis. Kluwer Academic Publishers, 1999.

[46] S. van Engeland, P.R. Snoeren, H. Huisman, C. Boetes, and N. Karssemeijer, 'Volumetric breast density estimation from full-field digital mammograms'. IEEE Transactions on Medical Imaging, 2006, 25, 273–282, doi:10.1109/TMI.2005.862741.

[47] R. Highnam, M. Brady, M.J. Yaffe, N. Karssemeijer, and J. Harvey, 'Robust breast composition measurement - Volpara'. In 'Digital Mammography', Berlin, Heidelberg, 2010 342–349, doi:10.1007/978-3-642-13666-5_46.

[48] K.R. Brandt, C.G. Scott, L. Ma, A.P. Mahmoudzadeh, M.R. Jensen, D.H. Whaley, F.F. Wu, S. Malkov, C.B. Hruska, A.D. Norman, J. Heine, J. Shepherd, V. S. Pankratz, K. Kerlikowske, and C.M. Vachon, 'Comparison of clinical and automated breast density measurements: Implications for risk prediction and supplemental screening'. Radiology, 2016, 279 (3), 710–719, doi:10.1148/radiol.2015151261.

[49] J.M. Seo, E.S. Ko, B.-K. Han, E.Y. Ko, J.H. Shin, and S.Y. Hahn, 'Automated volumetric breast density estimation: a comparison with visual assessment'. Clinical Radiology, 2013, 68 (7), 690–695, doi:10.1016/j.crad.2013.01.011.

[50] H.N. Lee, Y. Sohn, and K.H. Han, 'Comparison of mammographic density estimation by Volpara software with radiologists' visual assessment: analysis of clinical-radiologic factors affecting discrepancy between them'. Acta Radiologica, 2014, 56 (9), 1061–1068, doi:10.1177/0284185114554674.

[51] M.G.J. Kallenberg, C.H. van Gils, M. Lokate, G.J. den Heeten, and N. Karssemeijer, 'Effect of compression paddle tilt correction on volumetric breast density estimation'. Physics in Medicine and Biology, 2012, 57, 5155–5168, doi:10.1088/0031-9155/57/16/5155.

[52] J. Wang, A. Azziz, B. Fan, S. Malkov, C. Klifa, D. Newitt, S. Yitta, N. Hylton, K. Kerlikowske, and J.A. Shepherd, 'Agreement of mammographic measures of volumetric breast density to MRI'. PLOS ONE, 2013, 8, e81653, doi:10.1371/journal.pone.0081653.

[53] A. Gubern-Mérida, M. Kallenberg, B. Platel, R.M. Mann, R. Marti, and N. Karssemeijer, 'Volumetric breast density estimation from full-field digital mammograms: A validation study'. PLOS ONE, 2014, 9, e85952, doi:10.1371/journal.pone.0085952.

[54] W. He, A. Juette, E.R.E. Denton, A. Oliver, R. Martí, and R. Zwiggelaar, 'A review on automatic mammographic density and parenchymal segmentation'. International Journal of Breast Cancer, 2015, 2015, 1–31, doi:10.1155/2015/276217.

[55] J.W. Byng, N.F. Boyd, E. Fishell, R.A. Jong, and M.J. Yaffe, 'The quantitative analysis of mammographic densities'. Physics in Medicine and Biology, 1994, 39 (10), 1629–1638, doi:10.1088/0031-9155/39/10/008.

[56] S. Ciatto, N. Houssami, D. Bernardi, F. Caumo, M. Pellegrini, S. Brunelli, P. Tuttobene, P. Bricolo, C. Fantò, M. Valentini, S. Montemezzi, and P. Macaskill, 'Integration of 3D digital mammography with tomosynthesis for population breast-cancer screening (STORM): a prospective comparison study'. The Lancet Oncology, 2013, 14, 583–589, doi:10.1016/S1470-2045(13)70134-7.

[57] K. Lång, I. Andersson, A. Rosso, A. Tingberg, P. Timberg, and S. Zackrisson, 'Performance of one-view breast tomosynthesis as a stand-alone breast cancer screening modality: results from the Malmö Breast Tomosynthesis Screening Trial, a population-based study'. European Radiology, 2016, 26 (1), 184–190, doi:10.1007/s00330-015-3803-3.

[58] W.A. Berg, Z. Zhang, D. Lehrer, R.A. Jong, E.D. Pisano, R.G. Barr, M. Böhm-Vélez, M.C. Mahoney, W.P. Evans, 3rd, L.H. Larsen, M.J. Morton, E.B. Mendelson, D.M. Farria, J.B. Cormack, H.S. Marques, A. Adams, N.M. Yeh, G. Gabrielli, and A.C.R.I.N 6666 Investigators , 'Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk'. Journal of the American Medical Association, 2012, 307, 1394–1404, doi:10.1001/jama.2012.388.

[59] N. Ohuchi, A. Suzuki, T. Sobue, M. Kawai, S. Yamamoto, Y. Zheng, Y. Narikawa Shiono, H. Saito, S. Kuriyama, E. Tohno, T. Endo, A. Fukao, T. Tsuji, T. Yamaguchi, Y. Ohashi, M. Fukuda, T. Ishida, and for the J-START investigator group, 'Sensitivity and specificity of mammography and adjunctive ultrasonography to screen for breast cancer in the Japan Strategic Anti-cancer Randomized Trial (J-START): a randomised controlled trial'. The Lancet, 2016, 387 (10016), 341–348, doi:10.1016/S0140-6736(15)00774-6.

[60] B. Wilczek, H.E. Wilczek, L. Rasouliyan, and K. Leifland, 'Adding 3D automated breast ultrasound to mammography screening in women with heterogeneously and extremely dense breasts: Report from a hospital-based, high-volume, single-center breast cancer screening program'. European Journal of Radiology, 2016, 85 (9), 1554–1563, doi:10.1016/j.ejrad.2016.06.004.

[61] M.J. Emaus, M.F. Bakker, P.H.M. Peeters, C.E. Loo, R.M. Mann, M.D.F. de Jong, R.H.C. Bisschops, J. Veltman, K.M. Duvivier, M.B.I. Lobbes, R.M. Pijnappel, N. Karssemeijer, H.J. de Koning, M.A.A.J. van den Bosch, E.M. Monninkhof, W.P.Th.M. Mali, W.B. Veldhuis, and C.H. van Gils, 'MR Imaging as an additional

screening modality for the detection of breast cancer in women aged 50-75 years with extremely dense breasts: The DENSE trial study design'. Radiology, 2015, 277 (2), 527–537, doi:10.1148/radiol.2015141827.

[62] J. Melnikow, J.J. Fenton, E.P. Whitlock, D.L. Miglioretti, M.S. Weyrich, J.H. Thompson, and K. Shah, 'Supplemental screening for breast cancer in women with dense breasts: A systematic review for the U.S. Preventive Services Task Force'. Annals of Internal Medicine, 2016, 164 (4), 268–278, doi:10.7326/M15-1789.

[63] 'Are You Dense Advocacy Website'. http://www.areyoudenseadvocacy.org , accessed October 29, 2016.

[64] H.M. Gweon, J.H. Youk, J. Kim, and E.J. Son, 'Radiologist assessment of breast density by BI-RADS categories versus fully automated volumetric assessment'. American Journal of Roentgenology, 2013, 201 (3), 692–697, doi:10.2214/AJR.12.10197.

[65] A. Redondo, M. Comas, F. Macià, F. Ferrer, C. Murta-Nascimento, M.T. Maristany, E. Molins, M. Sala, and X. Castells, 'Inter- and intraradiologist variability in the BI-RADS assessment and breast density categories for screening mammograms'. British Journal of Radiology, 2012, 85 (1019), 1465–1470, doi:10.1259/bjr/21256379.

[66] A.M.J. Bluekens, N. Karssemeijer, D. Beijerinck, J.J.M. Deurenberg, R.E. van Engen, M.J.M. Broeders, and G.J. den Heeten, 'Consequences of digital mammography in population-based breast cancer screening: initial changes and long-term impact on referral rates'. European Radiology, 2010, 20 (9), 2067–2073, doi:10.1007/s00330-010-1786-7.

[67] N. Karssemeijer, A.M. Bluekens, D. Beijerinck, J.J. Deurenberg, M. Beekman, R. Visser, R. van Engen, A. Bartels-Kortland, and M.J. Broeders, 'Breast cancer screening results 5 years after introduction of digital mammography in a population-based screening program'. Radiology, 2009, 253, 353–358, doi:10.1148/radiol.2532090225.

[68] A.M.J. Bluekens, R. Holland, N. Karssemeijer, M.J.M. Broeders, and G.J. den Heeten, 'Comparison of digital screening mammography and screen-film mammography in the early detection of clinically relevant cancers: A multicenter study'. Radiology, 2012, 265 (3), 707–714, doi:10.1148/radiol.12111461.

[69] K. Kerlikowske, R.A. Hubbard, D.L. Miglioretti, B.M. Geller, B.C. Yankaskas, C.D. Lehman, S.H. Taplin, E.A. Sickles, and Breast Cancer Surveillance Consortium , 'Comparative effectiveness of digital versus film-screen mammography

in community practice in the United States'. Annals of Internal Medicine, 2011, 155 (8), 493–502, doi:10.7326/0003-4819-155-8-201110180-00005.

[70] E.D. Pisano, C. Gatsonis, E. Hendrick, M. Yaffe, J.K. Baum, S. Acharyya, E.F. Conant, L.L. Fajardo, L. Bassett, C. D'Orsi, R. Jong, M. Rebner, and Digital Mammographic Imaging Screening Trial (DMIST) Investigators Group, 'Diagnostic performance of digital versus film mammography for breast-cancer screening'. New England Journal of Medicine, 2005, 353 (17), 1773–1783, doi:10.1056/NEJMoa052911.

[71] N.M. Hambly, M.M. McNicholas, N. Phelan, G.C. Hargaden, A. O'Doherty, and F.L. Flanagan, 'Comparison of digital mammography and screen-film mammography in breast cancer screening: a review in the Irish breast screening program'. American Journal of Roentgenology, 2009, 193 (4), 1010–1018, doi:10.2214/AJR.08.2157.

[72] P.A. van Luijt, J. Fracheboud, E.A.M. Heijnsdijk, G.J. den Heeten, H.J. de Koning, and National Evaluation Team for Breast Cancer Screening in Netherlands Study Group (N.E.T.B), 'Nation-wide data on screening performance during the transition to digital mammography: Observations in 6 million screens'. European Journal of Cancer, 2013, 49 (16), 3517–3525, doi:10.1016/j.ejca.2013.06.020.

[73] L.M. Henderson, T. Benefield, S.J. Nyante, M.W. Marsh, M.A. Greenwood-Hickman, and B.F. Schroeder, 'Performance of digital screening mammography in a population-based cohort of black and white women'. Cancer Causes and Control, 2015, 26 (10), 1495–1499, doi:10.1007/s10552-015-0631-3.

[74] K. Kemp Jacobsen, E.S. O'Meara, D. Key, D. S M Buist, K. Kerlikowske, I. Vejborg, B.L. Sprague, E. Lynge, and M. von Euler-Chelpin, 'Comparing sensitivity and specificity of screening mammography in the United States and Denmark'. International Journal of Cancer, 2015, 137 (9), 2198–2207, doi:10.1002/ijc.29593.

[75] C.S. Lee, M. Bhargavan-Chatfield, E.S. Burnside, P. Nagy, and E.A. Sickles, 'The national mammography database: Preliminary data'. American Journal of Roentgenology, 2016, 206 (4), 883–890, doi:10.2214/AJR.15.14312.

[76] 'International Cancer Screening Network (ICSN) Website'. http://healthcaredelivery.cancer.gov/icsn/ , Accessed November 1, 2015.

[77] H.D. Nelson, E.S. O'Meara, K. Kerlikowske, S. Balch, and D. Miglioretti, 'Factors associated with rates of false-positive and false-negative results from digital mammography screening: An analysis of registry data'. Annals of Internal Medicine, 2016, 164 (4), 226–235, doi:10.7326/M15-0971.

[78] J. Kaufhold, J.A. Thomas, J.W. Eberhard, C.E. Galbo, and D.E. González Trotter, 'A calibration approach to glandular tissue composition estimation in digital mammography'. Medical Physics, 2002, 29 (8), 1867–1880, doi:10.1118/1.1493215.

[79] O. Alonzo-Proulx, R.A. Jong, and M.J. Yaffe, 'Volumetric breast density characteristics as determined from digital mammograms'. Physics in Medicine and Biology, 2012, 57 (22), 7443–7457, doi:10.1088/0031-9155/57/22/7443.

[80] O. Alonzo-Proulx, G.E. Mawdsley, J.T. Patrie, M.J. Yaffe, and J.A.Harvey, 'Reliability of automated breast density measurements'. Radiology, 2015, 275 (2), 366–376, doi:10.1148/radiol.15141686.

[81] E.S. Ko, R.B. Kim, and B. Han, 'Reproducibility of automated volumetric breast density assessment in short-term digital mammography reimaging'. Clinical Imaging, 2015, 39 (4), 582–586, doi:10.1016/j.clinimag.2015.02.011.

[82] R. Highnam and B. Schroeder, 'Asessing breast density change over time'. In '5th International workshop on breast densitometry and breast cancer risk assessment', 2011 38.

[83] S. Vanbelle and A. Albert, 'A bootstrap method for comparing correlated kappa coefficients'. Journal of Statistical Computation and Simulation, 2008, 78 (11), 1009–1015, doi:10.1080/00949650701410249.

[84] J.R. Landis and G.G. Koch, 'The measurement of observer agreement for categorical data'. Biometrics, 1977, 33 (1), 159–174, doi:10.2307/2529310.

[85] S. Vanbelle, 'Bootstrap program- pairwise agreement'. 2013, http://www.researchgate.net/publication/259810340_program-bootstrap_2013.

[86] L.T.W. de Jong-van den Berg, A. Faber, and P.B. van den Berg, 'HRT use in 2001 and 2004 in the Netherlands–a world of difference'. Maturitas, 2006, 54 (2), 193–197, doi:10.1016/j.maturitas.2005.10.010.

[87] K. Holland, M. Kallenberg, R. Mann, C. van Gils, and N. Karssemeijer, 'Stability of volumetric tissue composition measured in serial screening mammograms'. In Hiroshi Fujita, Takeshi Hara, and Chisako Muramatsu (eds.), 'Breast Imaging', volume 8539, 2014 239–244, doi:10.1007/978-3-319-07887-8_34.

[88] G. Gennaro and R. Highnam, 'This is what volumetric breast density is'. In 'European Congress of Radiology', 2013 doi:10.1594/ecr2013/C-1033.

[89] M.C. Spayne, C.C. Gard, J. Skelly, D.L. Miglioretti, P.M. Vacek, and B.M. Geller, 'Reproducibility of BI-RADS breast density measures among community radiologists: A prospective cohort study'. Breast Journal, 2012, 18, 326–333, doi:10.1111/j.1524-4741.2012.01250.x.

[90] J.A. Harvey, C.C. Gard, D.L. Miglioretti, B.C. Yankaskas, K. Kerlikowske, D.S.M. Buist, B.A. Geller, T.L. Onega, and Breast Cancer Surveillance Consortium, 'Reported mammographic density: film-screen versus digital acquisition'. Radiology, 2013, 266 (3), 752–758, doi:10.1148/radiol.12120221.

[91] J.M. Singh, E.M. Fallenberg, F. Diekmann, D.M. Renz, R. Witlandt, U. Bick, and F. Engelken, 'Volumetric breast density assessment: reproducibility in serial examinations and comparison with visual assessment'. RöFo- Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren, 2013, 185, 844–848, doi:10.1055/s-0033-1335981.

[92] J.O.P. Wanders, K. Holland, W.B. Veldhuis, R.M. Mann, R.M. Pijnappel, P.H.M. Peeters, C.H. van Gils, and N. Karssemeijer, 'Volumetric breast density affects performance of digital screening mammography'. Breast Cancer Research and Treatment, 2017, 162 (1), 95–103, doi:10.1007/s10549-016-4090-7.

[93] S. Saadatmand, E.J.T. Rutgers, R.A.E.M. Tollenaar, H.M. Zonderland, M.G.E.M. Ausems, K.B.M.I. Keymeulen, M.S. Schlooz-Vries, L.B. Koppert, E.A.M. Heijnsdijk, C. Seynaeve, C. Verhoef, J.C. Oosterwijk, I. Obdeijn, H.J. de Koning, and M.M.A. Tilanus-Linthorst, 'Breast density as indicator for the use of mammography or MRI to screen women with familial risk for breast cancer (FaMRIsc): a multicentre randomized controlled trial'. BMC Cancer, 2012, 12 (1), 440, doi:10.1186/1471-2407-12-440.

[94] D. Saslow, C. Boetes, W. Burke, S. Harms, M.O. Leach, C.D. Lehman, E. Morris, E. Pisano, M. Schnall, S. Sener, R.A. Smith, E. Warner, M. Yaffe, K.S. Andrews, C.A. Russell, and American Cancer Society Breast Cancer Advisory Group, 'American Cancer Society guidelines for breast screening with MRI as an adjunct to mammography'. CA: A Cancer Journal for Clinicians, 2007, 57 (2), 75–89, doi:10.3322/canjclin.57.2.75.

[95] R.M. Mann, C.K. Kuhl, K. Kinkel, and C. Boetes, 'Breast MRI: guidelines from the European Society of Breast Imaging'. European Radiology, 2008, 18, 1307–1318, doi:10.1007/s00330-008-0863-7.

[96] H. Li, M.L. Giger, O.I. Olopade, and L. Lan, 'Fractal analysis of mammographic parenchymal patterns in breast cancer risk assessment'. Academic Radiology, 2007, 14, 513–521, doi:10.1016/j.acra.2007.02.003.

[97] A. Torrent, A. Bardera, A. Oliver, J. Freixenet, I. Boada, M. Feixes, R. Marti, X. Llado, J. Pont, E. Perez, S. Pedraza, and J. Marti, 'Breast density segmentation: A comparison of clustering and region based techniques'. In 'Digital Mammography', Berlin, Heidelberg, 2008 9–16, doi:10.1007/978-3-540-70538-3_2.

[98] A. Oliver, X. Lladó, E. Pérez, J. Pont, E.R.E. Denton, J. Freixenet, and J. Martí, 'A statistical approach for breast density segmentation'. Journal of Digital Imaging, 2010, 23, 527–537, doi:10.1007/s10278-009-9217-5.

[99] S. Malkov, J. Wang, K. Kerlikowske, S.R. Cummings, and J. Shepherd, 'Single x-ray absorptiometry method for the quantitative mammographic measure of fibroglandular tissue volume'. Medical Physics, 2009, 36, 5525–5536, doi:10.1118/1.3253972.

[100] N. Karssemeijer, 'Automated classification of parenchymal patterns in mammograms'. Physics in Medicine and Biology, 1998, 43, 365–378, doi:10.1088/0031-9155/43/2/011.

[101] M.G. Kallenberg, M. Lokate, C.H. van Gils, and N. Karssemeijer, 'Automatic breast density segmentation: an integration of different approaches'. Physics in Medicine and Biology, 2011, 56, 2715–2729, doi:10.1088/0031-9155/56/9/005.

[102] P.R. Snoeren and N. Karssemeijer, 'Thickness correction of mammographic images by means of a global parameter model of the compressed breast'. IEEE Transactions on Medical Imaging, 2004, 23, 799–806, doi:10.1109/TMI.2004.827477.

[103] J.E. de Groot, M.J.M. Broeders, W. Branderhorst, G.J. den Heeten, and C.A. Grimbergen, 'Mammographic compression after breast conserving therapy: controlling pressure instead of force'. Medical Physics, 2014, 41 (2), 023501, doi:10.1118/1.4862512.

[104] A. Gubern-Mérida, M. Kallenberg, R.M. Mann, R. Marti, and N. Karssemeijer, 'Breast segmentation and density estimation in breast MRI: a fully automatic framework'. IEEE Journal of Biomedical Health Informatics, 2015, 19, 349–357, doi:10.1109/JBHI.2014.2311163.

[105] N.J. Tustison, B.B. Avants, P.A. Cook, Y. Zheng, A. Egan, P.A. Yushkevich, and J.C. Gee, 'N4ITK: Improved N3 bias correction'. IEEE Transactions on Medical Imaging, 2010, 29 (6), 1310–1320, doi:10.1109/TMI.2010.2046908.

[106] J. Nederend, L.E.M. Duijm, M.W.J. Louwman, J.W. Coebergh, R.M.H. Roumen, P.N. Lohle, J.A. Roukema, M.J.C.M. Rutten, L.N. van Steenbergen, M.F. Ernst,

F.H. Jansen, M.L. Plaisier, M.J.H.H. Hooijen, and A.C. Voogd, 'Impact of the transition from screen-film to digital screening mammography on interval cancer characteristics and treatment - a population based study from the Netherlands'. European Journal of Cancer, 2014, 50 (1), 31–39, doi:10.1016/j.ejca.2013.09.018.

[107] W.H. Kim, J.M. Chang, J. Lee, A.J. Chu, M. Seo, H.M. Gweon, H.R. Koo, S.H. Lee, N. Cho, M.S. Bae, S.U. Shin, S.E. Song, and W.K. Moon, 'Diagnostic performance of tomosynthesis and breast ultrasonography in women with dense breasts: a prospective comparison study'. Breast Cancer Research and Treatment, 2017, 162 (1), 85–94, doi:10.1007/s10549-017-4105-z.

[108] J. Fracheboud, H.J. de Koning, P.M. Beemsterboer, R. Boer, J.H. Hendriks, A.L. Verbeek, B.M. van Ineveld, A.E. de Bruyn, and P.J. van der Maas, 'Nation-wide breast cancer screening in the Netherlands: Results of initial and subsequent screening 1990-1995'. International Journal of Cancer, 1998, 75 (5), 694–698, doi:10.1002/(SICI)1097-0215(19980302)75:5<694::AID-IJC6>3.0.CO;2-U.

[109] J.D.M. Otten, N. Karssemeijer, J.H.C.L. Hendriks, J.H. Groenewoud, J. Fracheboud, A.L.M. Verbeek, H.J. de Koning, and R. Holland, 'Effect of recall rate on earlier screen detection of breast cancers based on the Dutch performance Indicators'. JNCI Journal of the National Cancer Institute, 2005, 97 (10), 748–754, doi:10.1093/jnci/dji131.

[110] R.J.P. Weber, R.M.G. van Bommel, M.W. Louwman, J. Nederend, A.C. Voogd, F.H. Jansen, V.C.G. Tjan-Heijnen, and Lucien E.M. Duijm, 'Characteristics and prognosis of interval cancers after biennial screen-film or full-field digital screening mammography'. Breast Cancer Research and Treatment, 2016, 158 (3), 471–483, doi:10.1007/s10549-016-3882-0.

[111] K. Holland, C.H. van Gils, J.O.P. Wanders, R.M. Mann, and N. Karssemeijer, 'Quantification of mammographic masking risk with volumetric breast density maps: How to select women for supplemental screening'. In Georgia D. Tourassi and Samuel G. Armato (eds.), 'Medical Imaging 2016: Computer-Aided Diagnosis', SPIE, 2016 doi:10.1117/12.2216810.

[112] J.G. Mainprize, O. Alonzo-Proulx, R.A. Jong, and M.J. Yaffe, 'Quantifying masking in clinical mammograms via local detectability of simulated lesions'. Medical Physics, 2016, 43 (3), 1249–1258, doi:10.1118/1.4941307.

[113] B. Chen, Y. Wang, X. Sun, W. Guo, M. Zhao, G. Cui, L. Hu, P. Li, Y. Ren, J. Feng, and J. Yu, 'Analysis of patient dose in full field digital mammography'. European Journal of Radiology, 2012, 81 (5), 868–872, doi:10.1016/j.ejrad.2011.02.027.

[114] J.J Heine, K. Cao, and J.A. Thomas, 'Effective radiation attenuation calibration for breast density: compression thickness influences and correction'. Biomedical Engineering Online, 2010, 9 (1), 73, doi:10.1186/1475-925X-9-73.

[115] D.B. Kopans, Breast Imaging. Lippincott Williams & Wilkins, 3rd edition, 2006.

[116] R.S. Saunders, Jr and E. Samei, 'The effect of breast compression on mass conspicuity in digital mammography'. Medical Physics, 2008, 35 (10), 4464–4473, doi:10.1118/1.2977600.

[117] W. Branderhorst, J.E. de Groot, R. Highnam, A. Chan, M. Böhm-Vélez, M.J.M. Broeders, G.J. den Heeten, and C.A. Grimbergen, 'Mammographic compression–a need for mechanical standardization'. European Journal of Radiology, 2015, 84 (4), 596–602, doi:10.1016/j.ejrad.2014.12.012.

[118] R.E. Hendrick, E.D. Pisano, A. Averbukh, C. Moran, E.A. Berns, M.J. Yaffe, B. Herman, S. Acharyya, and C. Gatsonis, 'Comparison of acquisition parameters and breast dose in digital mammography and screen-film mammography in the American College of Radiology Imaging Network Digital Mammographic Imaging Screening Trial'. American Journal of Roentgenology, 2010, 194 (2), 362–369, doi:10.2214/AJR.08.2114.

[119] D. O'Leary, T. Grand, and L. Rainford, 'Image quality and compression force: the forgotten link in optimisation of digital mammography?' Breast Cancer Research, 2011, 13 (1), P10, doi:10.1186/bcr2962.

[120] C.E. Mercer, P. Hogg, R. Lawson, J. Diffey, and E.R.E. Denton, 'Practitioner compression force variability in mammography: a preliminary study'. The British Journal of Radiology, 2013, 86 (1022), 20110596, doi:10.1259/bjr.20110596.

[121] C.E. Mercer, K. Szczepura, J. Kelly, S.R. Millington, E.R.E. Denton, B. Borgen, R. amd Hilton, and P. Hogg, 'A 6-year study of mammographic compression force: Practitioner variability within and between screening sites'. Radiography, 2015, 21 (1), 68–73, doi:10.1016/j.radi.2014.07.004.

[122] G.G. Waade, N. Moshina, S. Sæbuødegård, P. Hogg, and S. Hofvind, 'Compression forces used in the Norwegian breast cancer screening program'. The British Journal of Radiology, 2017, 90 (1071), 20160770, doi:10.1259/bjr.20160770.

[123] B. Davey, 'Pain during mammography: Possible risk factors and ways to alleviate pain'. Radiography, 2007, 13 (3), 229–234, doi:10.1016/j.radi.2006.03.001.

[124] J.R. Dullum, E.C. Lewis, and J.A. Mayer, 'Rates and correlates of discomfort associated with mammography'. Radiology, 2000, 214 (2), 547–552, doi:10.1148/radiology.214.2.r00fe23547.

[125] F.J. Keefe, E.R. Hauck, J. Egert, B. Rimer, and P. Kornguth, 'Mammography pain and discomfort: a cognitive-behavioral perspective'. Pain, 1994, 56 (3), 247–260, doi:10.1016/0304-3959(94)90163-5.

[126] P. Whelehan, A. Evans, M. Wells, and S. Macgillivray, 'The effect of mammography pain on repeat participation in breast cancer screening: A systematic review'. The Breast, 2013, 22 (4), 389–394, doi:10.1016/j.breast.2013.03.003.

[127] J.E. de Groot, W. Branderhorst, C.A. Grimbergen, G.J. den Heeten, and M.J.M. Broeders, 'Towards personalized compression in mammography: a comparison study between pressure- and force-standardization'. European Journal of Radiology, 2015, 84 (3), 384–391, doi:10.1016/j.ejrad.2014.12.005.

[128] J.E. de Groot, M.J.M. Broeders, W. Branderhorst, G.J. den Heeten, and C.A. Grimbergen, 'A novel approach to mammographic breast compression: Improved standardization and reduced discomfort by controlling pressure instead of force'. Medical Physics, 2013, 40 (8), 081901, doi:10.1118/1.4812418.

[129] N. Perry, M. Broeders, C. de Wolf, S. Törnberg, R. Holland, and L. Von Karsa, European Guidelines for Quality Assurance in Breast Cancer Screening and Diagnosis. 4 edition, 2008.

[130] R.J. Brenner, 'Asymmetric densities of the breast: Strategies for imaging evaluation'. Seminars in Roentgenology, 2001, 36 (3), 201–216, doi:10.1053/sroe.2001.25118.

[131] C.S. Giess, S.A. Chikarmane, D.A. Sippo, and R.L. Birdwell, 'Breast MR imaging for equivocal mammographic findings: Help or hindrance?' Radiographics, 2016, 36, 943–958, doi:10.1148/rg.2016150205.

[132] S.H. Heywang-Koebrunner, I. Schreer, and S. Barter, Diagnostic Breast Imaging: Mammography, Sonography, MRI and Interventional Procedures. Thieme Georg Verlag, Stuttgart, Germany, 3rd edition, 2014.

[133] K. Holland, I. Sechopoulos, G.J. den Heeten, R.M. Mann, and N. Karssemeijer, 'Performance of breast cancer screening depends on mammographic compression'. In A. Tingberg (ed.), 'Breast Imaging', volume 9699, 2016 183–189, doi:10.1007/978-3-319-41546-8_24.

[134] W. Branderhorst, J.E. de Groot, M. van Lier, R.P. Highnam, C.A. Grimbergen, and G.J. den Heeten, 'Validation of two methods of measuring contact area for estimation of applied compression pressure in mammography'. In 'Radiological Society of North America 2016 Scientific Assembly and Annual Meeting', 2016 .

[135] S. Hofvind, P.M. Vacek, J. Skelly, D.L. Weaver, and B.M. Geller, 'Comparing screening mammography for early breast cancer detection in Vermont and Norway.' JNCI Journal of the National Cancer Institute, 2008, 100 (15), 1082–1091, doi:10.1093/jnci/djn224.

[136] D.R. Busch, R. Choe, T. Durduran, D.H. Friedman, W.B. Baker, A.D. Maidment, M.A. Rosen, M.D. Schnall, and A.G. Yodh, 'Blood flow reduction in breast tissue due to mammographic compression'. Academic Radiology, 2014, 21 (2), 151–161, doi:10.1016/j.acra.2013.10.009.

[137] S.A. Carp, J. Selb, Q. Fang, R. Moore, D.B. Kopans, E. Rafferty, and D.A. Boas, 'Dynamic functional and mechanical response of breast tissue to compression'. Optics Express, 2008, 16 (20), 16064–16078, doi:10.1364/OE.16.016064.

[138] M. Dustler, I. Andersson, H. Brorson, P. Fröjd, S. Mattsson, A. Tingberg, S. Zackrisson, and D. Förnvik, 'Breast compression in mammography: pressure distribution patterns'. Acta Radiologica, 2012, 53 (9), 973–980, doi:10.1258/ar.2012.120238.

[139] W.A Berg, 'Current status of supplemental screening in dense breasts'. Journal of Clinical Oncology, 2016, 34 (16), 1840–1843, doi:10.1200/JCO.2015.65.8674.

[140] J.O.P. Wanders, K. Holland, N. Karssemeijer, P.H.M. Peeters, W.B. Veldhuis, R.M. Mann, and C.H. van Gils, 'The effect of volumetric breast density on the risk of screen detected and interval breast cancers: a cohort study'. Breast Cancer Research, 2017, 19 (1), 67, doi:10.1186/s13058-017-0859-9.

[141] J.O.P. Wanders, C.H. van Gils, N. Karssemeijer, K. Holland, M. Kallenberg, P.H.M. Peeters, M. Nielsen, and M. Lillholm, 'The combined effect of mammographic texture and density on breast cancer risk'. submitted, 2017.

[142] N.R. Kressin, C.M. Gunn, and T.A. Battaglia, 'Content, readability, and understandability of dense breast notifications by state'. Journal of the American Medical Association, 2016, 315 (16), 1786–1788, doi:10.1001/jama.2016.1712.

Acknowledgments

Recently, I read the following: “Wir sind alle auch Kinder des Zufalls, des Schicksals. Beeinflusst von Umfeld und Umständen und eben auch von unsere Herkunft. [...] Wir wurden und werden geprägt von Freunden und Klassenkameraden, von guten und schlechten Lehreren, Chefs und Kollegen, von Mut- und von Angstmachern, von unseren Eltern, die schon von ihren Eltern geprägt wurden”[1]. Which can be translated as: “We are all also children of chance, of destiny. Influenced by environment and circumstances and also by our origin. [...] We were and are shaped by friends and classmates, by good and bad teachers, bosses and colleagues, by courage and fear, by our parents, who were shaped by their parents.” It is true: There are many people, comments, situations and coincidences that have led to today, the day that I am writing these acknowledgments. Of course, it is not possible to thank everyone, but there are a few people who definitely deserve a thank you.

First of all, I would like to thank my promoter. Dear prof. dr. ir. Karssemeijer, dear Nico, this thesis would not have been possible without your in-depth knowledge about breast cancer, mammography and breast density. In our weekly meetings, you came up with remarkable ideas: sometimes simple ones which made me wonder why I did not think of them myself, but also challenging ones which I had to ponder on for some time. Thank you for your inspiring thoughts and efforts to help complete the thesis; I learned a lot from you.

My thanks also go to my co-promoters. Though we did not meet regularly, I would like to thank you for your input during all stages of my research. Dear dr. van Gils, dear Carla, your epidemiological point of view was always welcome and helpful. Our discussions made me confident that we were using the correct statistical tests and your comments on my writing made sure to stay focused on the (statistical) essentials. Dear dr. Mann, dear Ritse, I would like to thank you for your input from the clinic. You were always available to answer all my questions regarding breast imaging and clinical work flow, regardless of how minor or trivial they were. Last, but definitely not least, thank you for the endless hours spent on the scoring of breast density.

Prof. dr. Verbeek, prof. dr. Pijnapple and dr. Lillholm, I would like to thank you for participating in the doctoral thesis committee.

My thanks also go to the co-authors of the papers and abstracts. This thesis would look completly different without your density readings and contributions. Furthermore, I would like to thank everyone who has read and commented on the introduction or summaries.

My research was embedded into the ASSURE project; I would like to thank all the col-

[1] Henning Sussebach, “Es ist denkbar, dass man mit Offenheit mehr erreicht als mit Verdruckstheit”, Die Zeit N°20

laborators for the fruitful discussions.

This research would not have been possible without data, a thanks goes to the population screening Mid-West for the images and the LBO for the great collaboration and the efforts regarding meta data.

Mads Nielsen and Martin Lillholm: thank you for the pleasant and inspiring time at DIKU and Biomediq in København. Though this period did not lead to a publication, I had a wonderful summer.

Dear Hanneke, we have been working on the same dataset for years now and our complimentary research has led to many new ideas and insights. It was really helpful to look at the same data from different points of view. Unfortunately, we communicated mostly via email; personal meetings were limited to a few times per year. Nevertheless, we managed to go to one conference together. Our bike tour through San Francisco with Carla, Marije and Stéphanie was definitely the ideal start to my vacation in the US west coast in 2014. I wish you all the best for your future career.

Next, I want to thank everyone I met at DIAG. DIAG has always been a nice working place and a friendly environment. I enjoyed being in the office and out on DIAG-weekends. Thank you for the great discussions; I learned a lot. Starting at "tropical island -1", I had several offices and room mates, the following people should not go unnoticed: Albert, Babak, Colin, Freerk, Isabell, Jan, Jan-Jurre, Kaman, Miriam, Mohsen, Paul, Pragnya, Sjoerd, Tom and Rick.

Suzan, as a member of "kamer 25"you should be on the previous list. Instead, together with Christiana, you deserve a special thanks: Thank you for being my paranymphs and thank you for the useful and less useful discussions we had. I wish you all the best in finishing your theses, I will be more than happy to come back to Nijmegen to attend your defences.

Now, I would like to go a little bit back in time. A study was necessary to start a PhD. During this study, I had the possibility to organise a study tour. Though I liked the idea of organising a tour when I read the corresponding Marie Curie newsletter for the first time, I hesitated to volunteer as I was hardly involved with the student association. Guus, Maaike and Remko thank you for asking me to be part of the McReis-team; I never regretted this decision and this journey through the US is one of the best memories of my student days. Vincent and Wouter, after spending many hours in the same lectures, I got to know you while being at CERN and at our gatherings afterwards. I have to admit that I could hardly

follow your discussions for more than two minutes. Theoretical high energy physics is your passion, not mine. Fortunately, Amber was there as well and we had a great time together anyway. Thanks for introducing me to some Dutch traditions; going to the bevrijdingsfestival and marching in Wageningen are not on the typical German to-do list.

It is time to switch languages. Elisabeth und Marie, am Anfang einte uns vor allem eins, unsere Studienwahl und der Entschluss in Nijmegen zu studieren. In den letzten Jahren sind so viele Erinnerungen dazugekommen. Neben den Treffen in Nijmegen, gab es ja noch die Studienreise und Genf. Mein Studium wäre ohne euch ein anderes gewesen. Schade, dass wir uns so selten sehen und unsere Lebensmittelpunkte soweit auseinander liegen. Elisabeth, kannst du dir vorstellen das wir uns nun schon 10 Jahre kennen? Ich weiß noch, dass wir uns vor dem Bahnhof während der Stadtreally der Intro zum ersten Mal unterhielten. Die Zeit vergeht so rasend schnell. Marie, du lebst den Traum den wir alle mal hatten, Wissenschaftlerin am CERN zu sein. Ich wünsch euch viel Erfolg für die Zukunft.
Rachel und Nasti, auch wenn zwischenzeitlich mal Funkstille war, es ist schön, dass wir uns wieder, mehr oder weniger, regelmäßig treffen. Auch wenn unser Alltag sehr unterschiedlich ist, finden wir doch immer wieder Schnittstellen. Dazu kommt dann noch der Tratsch und Klatsch über ”alte” Zeiten.

Last but not least, auch einen Dank an meine Familie. Ich weiß nicht wie oft ich beim Samstags-/Sonntagskaffee gefragt wurde ”Was macht der Doktor?” und die Antwort ”läuft” oder ”wir arbeiten dran” war. Nun habt ihr das Resultat von 4 Jahren Arbeit in den Händen. Barbara, Schwesterherz, auch wenn wir in vielen Sachen grundverschieden sind, so habe wir auch viel gemeinsam. Oft brauche wir einander nur anschauen um zu wissen, dass wir das Gleiche denken. Ich bin davon überzeugt, dass die Masterarbeit die du gerade schreibst ein Erfolg wird und du deinen Weg gehen wirst. Schön, dass wir uns immer noch so oft sehen.
Mama, Papa, ohne euch wäre ich nicht wo ich jetzt bin. Vielen Dank für die Unterstützung und das Vertrauen, das ihr mir entgegengebracht habt. Ich weiß, ich kann immer auf euch zählen. Danke!

Curriculum Vitae

Katharina Holland was born in Goch (Germany) on February 4th 1988. In 2007, after graduating from high school (Gesamtschule Mittelkreis Goch) she started studying physics at the Radboud University in Nijmegen. She obtained her Bachelor of Science degree in 2011, followed by the Master of Science degree in 2012. In 2011, she spent six weeks at CERN, where she discovered her interests in medical physics when following lectures about accelerators for cancer treatment. Katharina joined the Diagnostic Image Analysis Group as a PhD student in November 2012. Under the supervision of prof. Nico Karssemeijer, dr. Carla van Gils and dr. Ritse Mann, she has been working on methods to quantify breast density and masking risk in mammographic images. The results of her research are described in this thesis.